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 automated registration of multi-modal sensor data. The metric is based on the alignment of the orientation of gradients formed from the two candidate sensors. Data registration is performed by estimating the sensors' extrinsic parameters that minimises the misalignment of the gradients. The metric can operate in a large range of applications working on both 2D and 3D sensor outputs and is suitable for both (i) single scan data registration and (ii) multi-sensor platform calibration using multiple scans. Unlike traditional calibration methods, it does not require markers or other registration aids to be placed in the scene. The effectiveness of the new method is demonstrated with experimental results on a variety of camera-lidar and camera-camera calibration problems. The novel metric is validated through comparisons with state of the art methods. Our approach is shown to give high quality registrations under all tested conditions.

1 Introduction

One of the most difficult problems in building multi-modal and/or multi-temporal representations of the environment is accurate data registration. In multi-modal mapping, the information provided by two sensors can be integrated by estimating the sensors' relative location and orientation. For applications requiring high precision results, this location cannot be found by simply measuring the sensor's relative positions due to the uncertainty introduced in the measurement process and the uncertainty of the actual sensor location inside its casing. This registration is far from trivial due to the very different modalities via which the two sensors may operate (Le and Ng, 2009). In mobile robotics where the sensors are typically rigidly mounted in a frame, calibration 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. A large number of methods have been developed using these techniques to calibrate multi-camera (Lébraly and Royer, 2011; Bouguet, 2004; Kumar and Ilie, 2008) and lidar-camera (Unnikrishnan and Hebert, 2005; Zhang and Pless, 2004; Bouguet, 2004) systems.

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. Note that for building multi-modal environment representation, the system does not require the ability to work in real-time, but rather is required to be able to estimate the calibration by processing a small subset of the data during the robot's regular operation.

This paper presents 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. For example, our approach is appropriate for single scan applications, such as registration of multi-temporal data or as shown in our experiments, to register data collected from non-rigidly mounted sensors. One of the applications that motivated the development of this metric was to align images and lidar scans taken at a mine site. The camera and lidar scanner were operated by two different teams, at different times and the only location provided was through use of a handheld GPS system. 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.

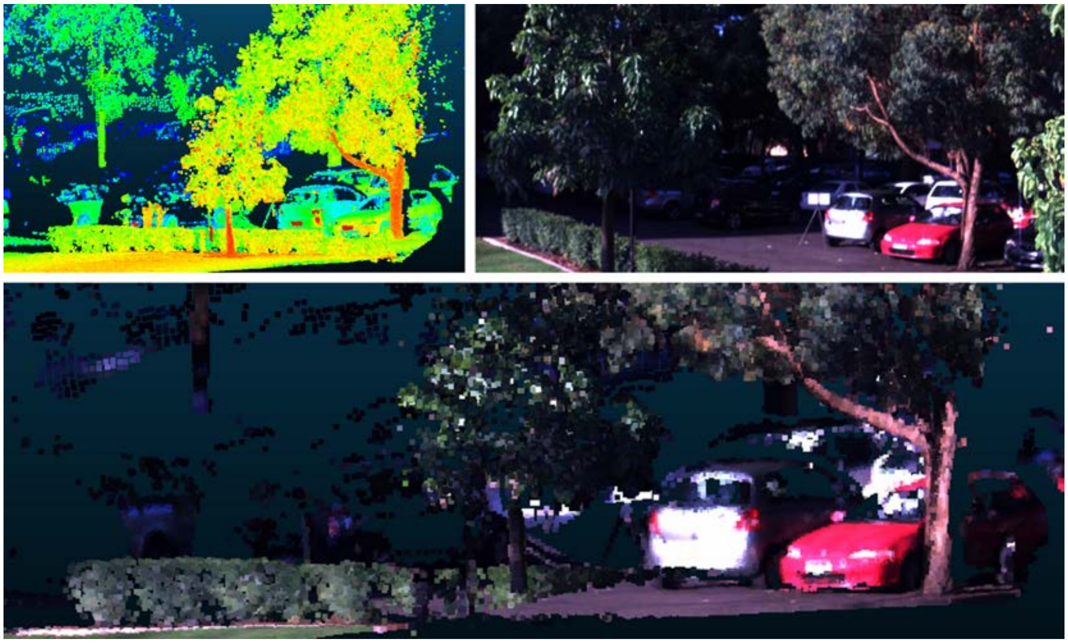


Figure 1: Camera and lidar scan being combined. Raw lidar data is shown in the top left, with the camera image shown in the top right. The textured map obtained with our approach is shown at the bottom.

Specifically, this paper presents the following contributions:

- A review of the state of the art in multi-modal data registration and markerless sensor calibration.
- The introduction of a new metric for automated registration and calibration of multi-modal data.
- The evaluation of the proposed approach in three different datasets, including different sensor modalities and different types of environment.
- A thorough comparison of our approach and three different techniques: Levinson and Thrun's method (Levinson and Thrun, 2012), mutual information as used by Pandey et al. (Pandey et al., 2012) and a normalised mutual information based method (Taylor and Nieto, 2012).

- The evaluation of different 3D features for use with the method proposed.

2 Related work

A large body of work exists for sensor calibration. To provide an in-depth review of the methods most relevant to our proposed technique, we have limited the scope of this section to methods that are both multimodal and markerless. To highlight the similarities and differences among these methods we will divide the related work into four sections. First we present a description of mutual information, followed by 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 is a measure of statistical dependency between two signals known as *mutual information* (MI). 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 methods presented in all three of the following sections of this literature review and 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), which is a measure of how much information is contained in a signal. Its discrete version is defined as:

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

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

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

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

When used in registration 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 *normalised mutual information* (NMI) which is defined as:

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

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 Multimodal 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 *medical resonance imaging* (MRI) and *computed tomography* (CT) scans for use in medical imaging.

Bodensteiner et al. successfully matched images using the mono-modal feature descriptor called the *scale invariant feature transform* (SIFT) (Bodensteiner et al., 2011). In most multi-modal applications, however, it is found that the assumptions made by SIFT about the directions and relative magnitudes of the gradients 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).

The most common methods used in medical image registration are the MI and NMI methods that have already been discussed. However, many other methods exist. The correlation ratio used by A. Rouche et al. provides a measure of the functional dependence of the intensities (Roche et al., 1998). A second closely related metric is the correlation coefficient; this is a less general metric providing a measure of the linear dependence of the two images intensities. A method known as gradient correlation takes the normalised cross-correlation between two gradient images created using a Sobel filter (Penney et al., 1998). Zana and Klein used a Hough transform to register retinal images; the vessels in the eye provided strong lines for this method to detect and align (Zana and Klein, 1999). Wachinger and Navab developed a method called *entropy sum of squared differences* (eSSD) that uses the entropy of patches of the images for registration of T1 and T2 MRI scans. The method works by first creating images where each pixel's intensity is equal to the entropy of the pixels in an n-by-n patch around it. Matching is then performed by taking the *sum of squared differences* (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. 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

These systems operate by matching a single high resolution scan of an environment with its corresponding image. High resolution scans are usually produced with 2D lidar scanners and can take up to several minutes to integrate all of the data into a single scan.

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 images of the same area was 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 camera's 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 this method worked well for the application presented, 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 operates by creating an accurate camera model that emulates the hyperspectral camera and using it to project the lidar points onto the camera's images. Some of the lidar point clouds did not have usable intensity information, therefore a new intensity was assigned to the points based on an estimation of the direction of their surface normals. These lidar points were compared to the points in the image that they were projected onto using NMI. It was assumed that 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 lidar's 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.

A more theoretical view on calibration is presented in (Corsini et al., 2009) where the authors looked into different techniques for generating a synthetic image from a 3D 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 combinations of the silhouette, normals, specular map and ambient occlusion 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

While in theory a similar problem to the single lidar scan case explored above, mobile systems usually have a large number of much lower resolution scans. This difference means that most methods developed for the above systems give poor results on these datasets. As the wide spread availability of 3D lidars capable of operating from a moving platform did not happen until the Velodyne HDL-64E was released in 2007, 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 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 have been independently developed by Pandey et al. (Pandey et al., 2012) and Wang et al. (Wang et al., 2012). 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.

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

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 facilitates comparisons of different alignments during the optimisation stage. The initial step is to assign an intensity value to each data point. For 2D data the average of the colour channels is used. For 3D data, the user selects one of several possible features usually depending on the exact sensor and application (the features considered will be presented in section 3.1). After this is done, 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; the edge detector used also 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. For the optimisation, our approach uses particle swarm for one-off data registration, and Nelder Mead simplex for multi-sensor platform calibration. Different optimisers are used as, in the platform calibration case, scans can be aggregated. This smooths the search space and allows a local optimiser to be used. [1]

3.1 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 3D range sensors, several possible features exist that can be used to set the points intensities. The features analysed in this work are: (i) the normals of the points, (ii) the return intensity and (iii) the distance of points from the sensors. Histogram equalisation is performed on all features to improve contrast.

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

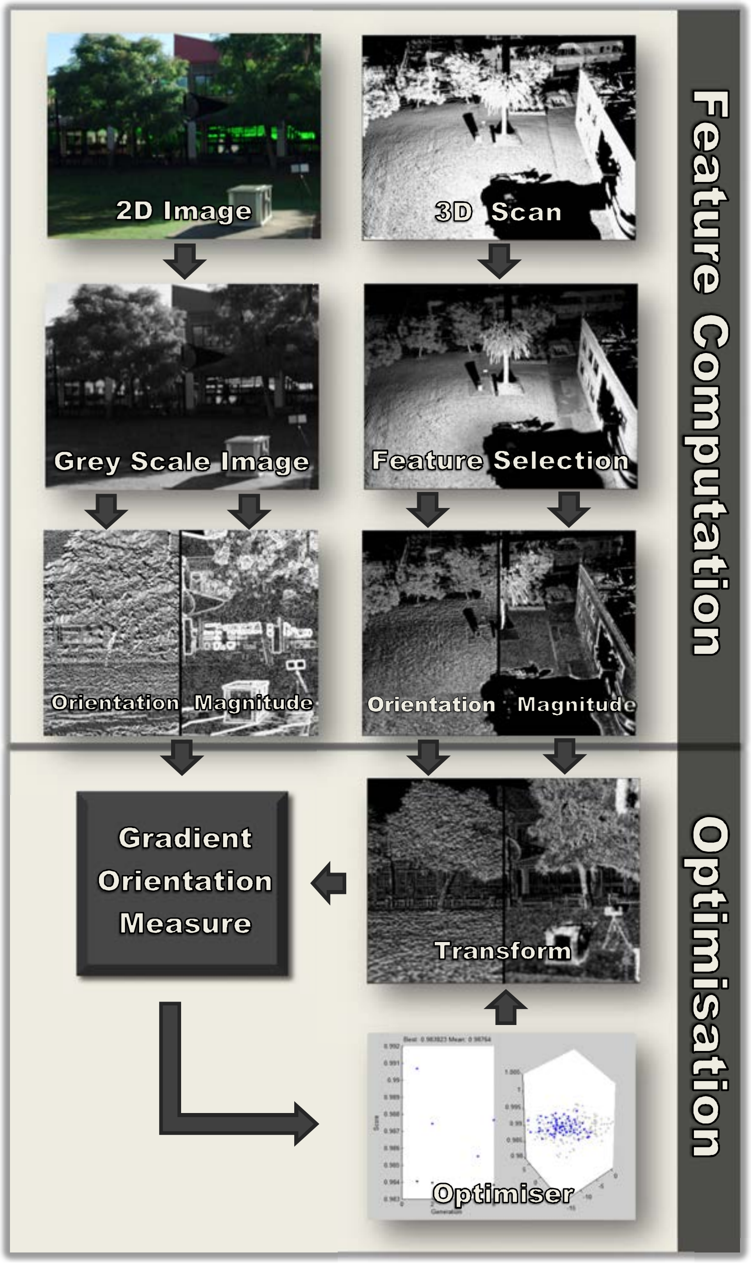


Figure 2: An example of the steps of our approach. This diagram shows the alignment of a camera image with a high resolution lidar scan coloured by its return intensity.

Normals of points: To obtain 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 which 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 eigenvector corresponding to the smallest eigenvalue of C is the best estimate of its normal vector to the plane. Once the normals have been calculated, the three values that make up the normal vector 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 intensity readings cannot be used. A second issue with this method occurs when using systems that make use of multiple lasers, such as the Velodyne scanner. In these systems, each laser scans a different section of the environment. For example, in the Velodyne, the scan is built up by rotating its head containing 64 separate lasers. Each of these lasers has slightly different characteristics and can give substantially different intensity of returns for the same object.

Range: 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 due to a lack of sharp changes in depth.

3.2 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.

3.2.1 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. In regular cameras, an image is created when light strikes a 2D *charge-coupled device* (CCD). However, in this panoramic camera, the CCD is a single vertical line array mounted on top of a motor and slowly rotated to build up a panoramic image of the environment. 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)

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

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

where
$\mathbf{x}_{cam}$, $\mathbf{y}_{cam}$ 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.

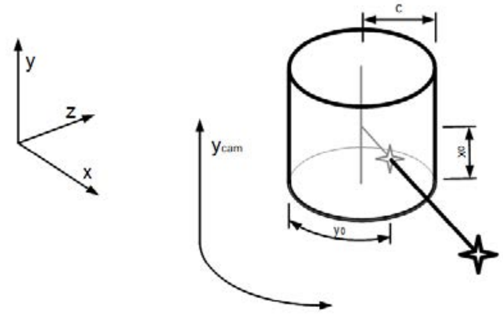


Figure 3: Cylinder model used to represent a panoramic camera.

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

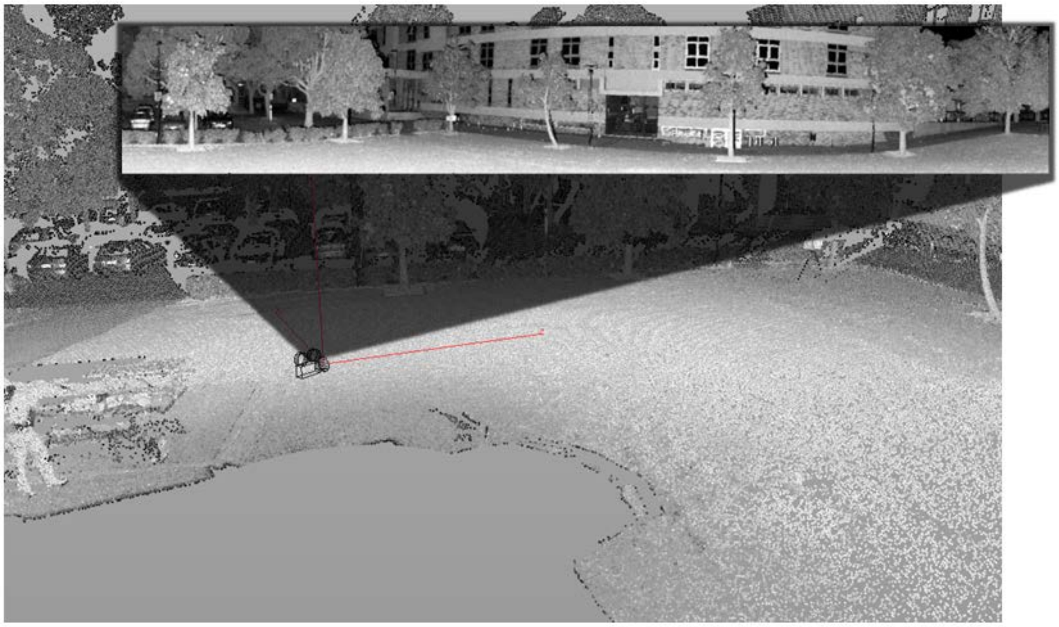


Figure 4: An image of a lidar scan. A virtual camera has been placed in the scan and is generating an image from the scan that will later be aligned with a real image of the same scene.

3.2.2 Affine transform

To perfectly align all points in two images taken by two cameras at different locations, the distance to each point in the images must be known. This is due to parallax changing where objects appear in each image. This alignment can be estimated using stereo vision methods. However, the correspondence between pixels must be recalculated for each frame. Due to the difficulty of performing such alignment on two cameras of different modality, we instead calibrate the camera images using a simple affine transform. While not perfect, for two cameras with only small differences in location and orientation, an affine transform can give high quality image registration. An affine transform was chosen over a projective transform to reduce the number of parameters required to optimise.

3.3 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. The gradient of the points in the 3D data should be from the perspective of the camera so that, once transformed to the camera's frame of reference, the orientations for both sensors will be aligned. The most simple way to calculate the gradients for the 3D data is to use the camera model to generate an image and apply the standard Sobel operator to it. Unfortunately, this method gives poor results due to the problems outlined in Section 3.5. Instead, the gradient of the 3D points are calculated as described below.

The points are first projected onto a sphere that is centred at the estimated location of the camera using Equations 7 and 8. A sphere is used rather than a plane as in image generation, since with a plane, points in front of and behind the camera can adversely be projected onto the same location. Each point on the sphere has its 8 nearest neighbours evaluated before the gradient is calculated by adding the gradient vectors to each neighbouring point. The process for this is shown in Algorithm 1.

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

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

```
Let
$p_x, p_y, p_v$ be the current point's 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
$g, temp, g_{mag}, g_{or}$ be the gradient vector, a temporary vector, the gradient vector's magnitude and the gradient vector's orientation, respectively

$g = 0$;
for neighbouring point $n$ do
    $x = p_x - n_x$;
    $y = p_y - n_y$;
    $v = p_v - n_v$;
    $temp_{mag} = \frac{v}{8*\sqrt{x^2+y^2}}$;
    $temp_{or} = arctan2(y, x)$;
    $g = g + temp$;
end
```

Algorithm 1: Gradient calculation for 3D sensors

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 computationally 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 great reduction in the computational cost. We concluded that this assumption would be valid for most practical cases as the search space used for optimising the lidar's extrinsic calibration is usually most 1 m and 5 degrees, whereas the distance to most objects in the environment in all our experiments was well over 10 m. This meant that there would have been only minor changes in the calculated gradients' magnitude and orientation. To test the validity of this, a simple experiment was run on one of our datasets. A lidar scan of the Australian Centre for Field Robotics building (ACFR scan 1 from dataset 1 presented in Section 4.4) was first aligned using GOM without the simplifying assumption of constant gradient values. The gradients were then recalculated for different levels of offset from the position and perspective of the camera. These scans were used to calculate the GOM values over a range of camera yaw values. The results are shown in Figure 5.

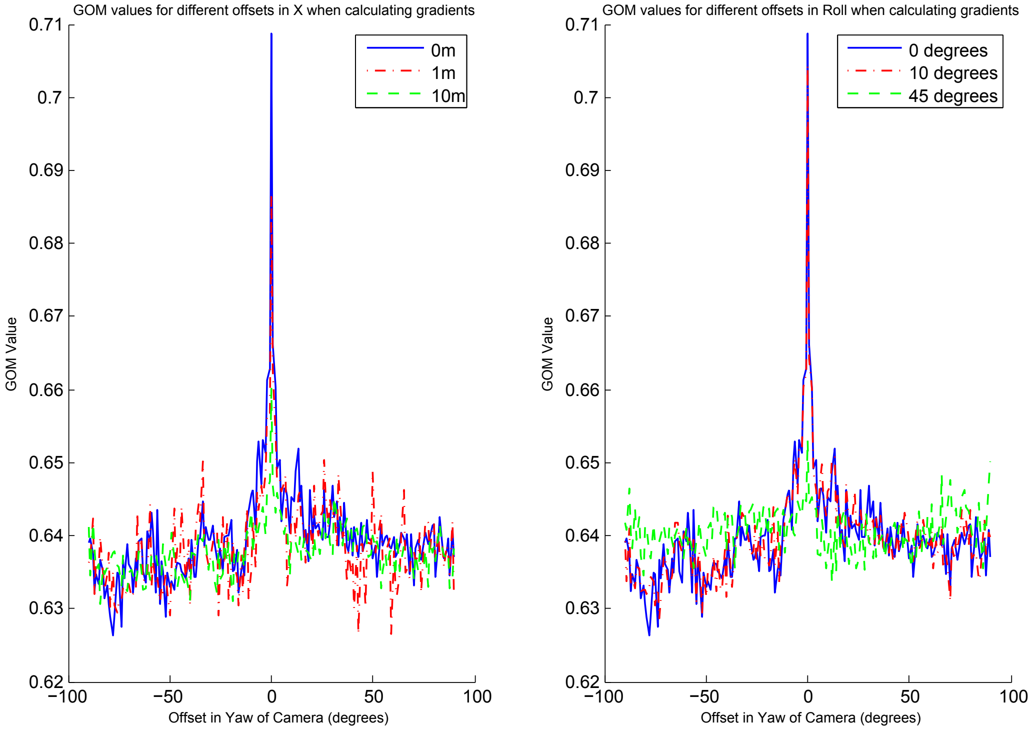


Figure 5: The plots illustrate how offsets in the position and orientation of the lidar affect GOM, when assuming constant gradient values. GOM is plotted for a range of yaw values to show the global maximum for each run.

Three different levels of offset were introduced into the X location the gradients were calculated at. These offsets were 0, 1 and 10 m. Despite the different locations where the gradients were calculated, the global maximum for GOM still occurred in the same location for all three runs. While the value of the global maximum for the 1 and 10 m error runs was slightly lower than that of the 0 m run, it was still clearly distinct from other local maxima. Similar results were also obtained from different datasets and introducing errors into the Y and Z dimension.

The same experiment was performed for the orientation of the lidar. For its roll, three different errors of 0, 10 and 45 degrees were used. A roll offset of 10 degrees has little impact on the results. However, a roll offset of 45 degrees significantly reduced the value of the global maximum. This is expected as a change in roll has the most direct impact on the orientation of the gradients, and therefore an initial offset as large as

45 degrees breaks the assumption that $parameters_{initial} \approx parameters_{final}$. Pitch was found to have less impact, and yaw is independent of the gradients. This experiment showed that for our application, while this assumption would slightly degrade the value of the global maximum, it would, be unlikely to shift it significantly, making the large reduction in computational time offered by the assumption worthwhile.

3.4 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. This correspondence even exists between seemingly unrelated features such as range and reflectance. For example, a sharp change in distance usually indicates a change in the object being detected. There is a high probability that these objects will be made of materials with different reflectance properties, meaning that it is likely that there will be a significant change in reflectance at the same location.

GOM exploits these differences to give a measure of the alignment. GOM was inspired by a measure proposed in (Pluim et al., 2000) for use in medical imaging registration. The presented measure, however, has several differences as Pluim J. et al's. method is un-normalised; uses a different calculation of the gradient's strength and direction, 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 the points are by taking the absolute value of the dot product of the gradient vectors:

$$alignment_j = |g_{(1,j)} \cdot g_{(2,j)}| \tag{9}$$

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. This means that for two aligned images, the two corresponding gradients may be out of phase by 180 degrees. An example of this occurring is shown in Figure 6.

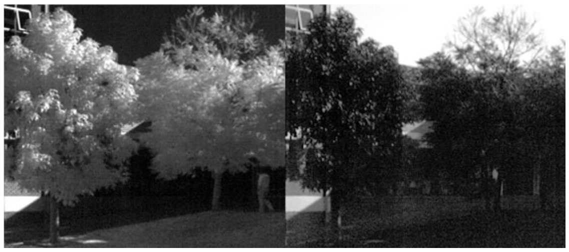


Figure 6: A pair of example images showing how gradients may reverse direction between modalities. In the left image, taken at 950 nm, the trees are white with a black sky behind them. However in the right image, taken at 418 nm, the trees appear black and the sky white. This means that if the gradient between the trees and sky are calculated, the gradients will be 180 degrees out of phase for the two images.

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. While this issue could be corrected by normalising the vectors before taking the dot product, 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, which normalising at this stage would remove. 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 10.

$$GOM = \frac{\sum_{j=1}^{n} |g_{(1,j)} \cdot g_{(2,j)}|}{\sum_{j=1}^{n} \|g_{(1,j)}\| \|g_{(2,j)}\|} \tag{10}$$

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 7. The NMI values are also shown for comparison.

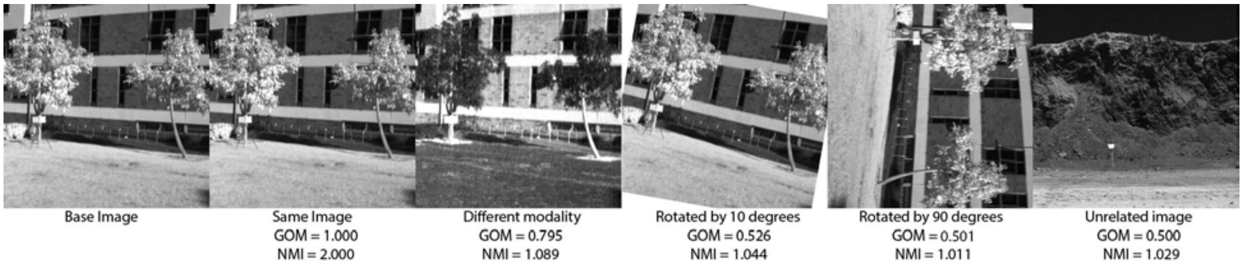


Figure 7: GOM and NMI values when the base image shown on the left is compared with a range of other images.

3.5 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) cause the majority of the pixels to 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 methods based on edge directions, such as GOM. A typical image produced from Velodyne data is shown in Figure 8. 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 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, linear interpolation is performed at the coordinates given by the point list.

3.6 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, as outlined below.

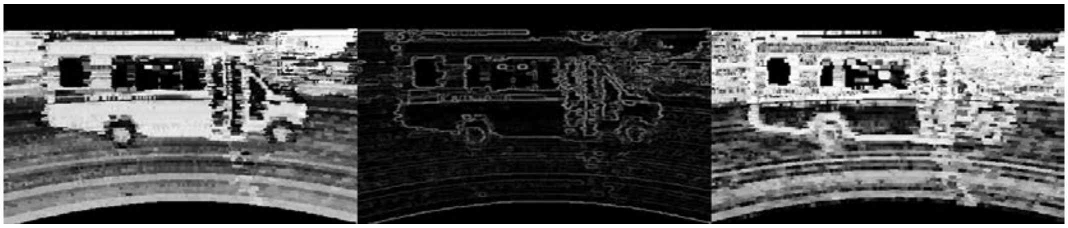


Figure 8: An Image generated from Velodyne lidar data is shown on the left. The centre image shows the image's gradients calculated using a standard Sobel filter, the right image shows the gradients calculated from the point cloud by our own technique. Due to the sparse nature of the lidar points when a Sobel filter is used, its gradients tend to depend more on the distribution of points than their intensity. In this instance this results in almost all the gradients being detected as horizontal or vertical lines.

3.6.1 Multi-sensor calibration

When multiple scans can be aggregated, as in the case of multi-sensor platforms calibration, the optimisation is performed using the Nelder-Mead simplex method. This method is a well established local optimisation technique (Nelder and Mead, 1965). It creates a simplex on the function and uses the points of this simplex to estimate the position of the maximum. It then uses this information to replace one of the points in the simplex. This process repeats until the function converges to a solution. For this optimiser to reliably converge, the search space must be convex. If this condition is not met, the optimisation can become trapped at a local maximum. Generally, this requirement would not be satisfied by any of the metrics analysed (GOM, MI ,NMI or the metric presented by Levinson et al.). However, for the case of multi-sensor calibration, it is a reasonable assumption for three key reasons:

1) In calibrating a multisensor system, an accurate initial guess to the alignment of the sensors is usually known. This is because the sensors are mounted in a rigid setup with the only unknown being the location of the sensors within their housings.

2) As the sensors are rigidly mounted, a large number of sensor readings can be aggregated. Aggregating readings is beneficial as it both provides additional information and helps to minimise the impact of alignments between unrelated areas. The effect of this is to reduce the strength and number of local maxima produced by the metric.

3) The basin of attraction for GOM's global maximum can be substantially increased by applying a Gaussian blur to the output of one of the sensors it is aligning. This works as the GOM makes use of overlapping edges to calculate its values, and unless the edges are partially overlapping, there is nothing to indicate how close to the correct alignment the measure is. When blurring is applied, the size of the edges is increased, providing more overlap between edges, and an improved indication of the alignment. An issue with this blurring is that it removes many of the small edges that can be used to indicate a precise alignment, reducing the accuracy of the metric. To prevent this issue, an optimisation pyramid was implemented where the result of a level of the pyramid is used as the initial guess for the optimisation of the same image with less blurring. In our experiments, four layers were used, with Gaussians with σ of 4, 2, 1 and 0 applied. The effect of the aggregation and blurring can be seen in Figure 9.

3.6.2 Registration

Optimisation of the registration problem with single scans is generally significantly more challenging than the calibration problem. The main reason for this is that the error in the initial guess for the sensors' alignments can be very large. In many situations the initial conditions are given using consumer GPS systems or a simple note claiming that the sensors were positioned close to each other. This can lead to initial position errors in excess of 1 m and 10 degrees. There is also no rigid connection between the sensors, preventing

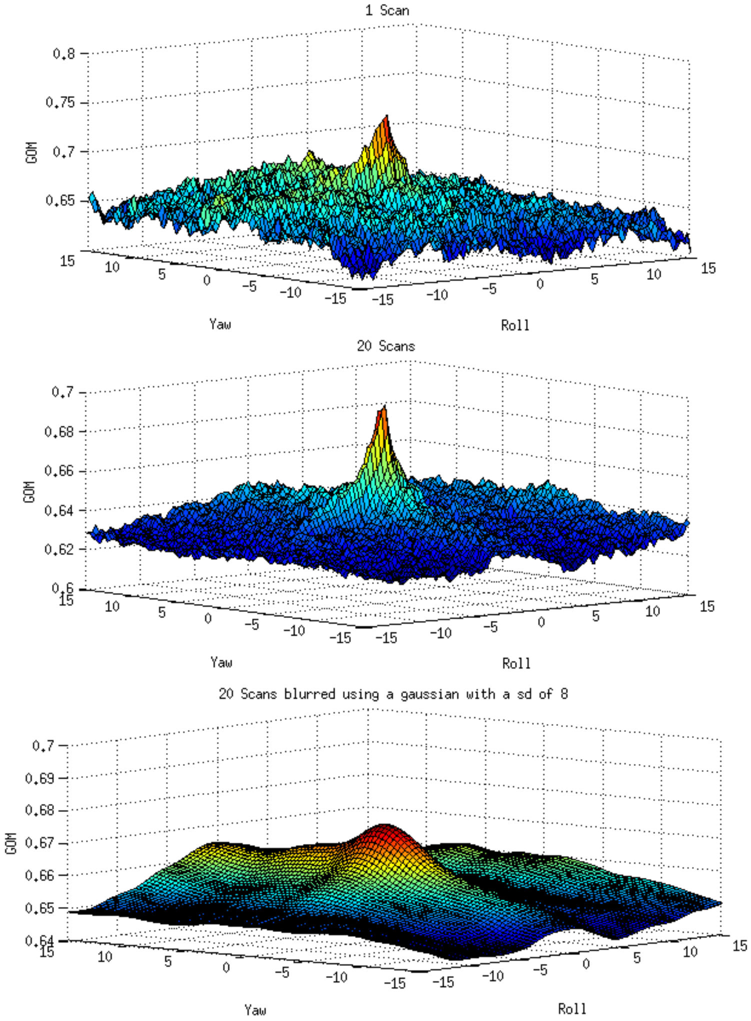


Figure 9: GOM values plotted for the roll and yaw of a typical lidar-photo alignment using one scan-image pair (top) and 20 scan-image pairs (middle). Also shown is the result of applying a Gaussian blur to the 20 images. (bottom). Aggregation and Gaussian blurring both significantly smooth the function, allowing for easier optimisation of the metric

scan aggregation from being used. These limitations result in a highly non-convex search space that requires a global optimisation technique to find the maximum.

We evaluated several different global optimisation methods such as pattern search (Audet and Jr, 2002), global search (Ugray and Lasdon, 2007), genetic algorithms (Goldberg and Holland, 1988), exhaustive search and particle swarm optimisation (Mikki and Kishk, 2008). Particle swarm was found to perform the fastest of the options while still being robust. Its operation was also fairly intuitive, allowing the ability to judge how the optimisation is progressing as it runs.

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 process of registration is shown in Algorithm 2. The implementation of particle swarm used was developed by S Chen (Chen, 2009). In our experiments we used a particle swarm optimiser with 500 particles.

```
Let
$\boldsymbol{r^i(t)}$ be the position of particle i at time t
$\boldsymbol{v^i(t)}$ be the velocity of particle i at time t
$\boldsymbol{p_n^{i,L}}$ be the local best of the ith particle for the nth dimension
$\boldsymbol{p_n^g}$ be the global best for the nth dimension
$\boldsymbol{n \in 1, 2, ...N}$
$\boldsymbol{t}$ be the time
$\boldsymbol{\Delta t}$ be the time step
$\boldsymbol{c_1}$ and $\boldsymbol{c_2}$ are the cognitive and social factor constants
$\boldsymbol{\phi_1}$ and $\boldsymbol{\phi_2}$ are two statistically independent random variables uniformly distributed between 0 and 1
$\boldsymbol{w}$ be the inertial factor

for each iteration $l$ do
    if $f(r^i(l+1)) > f(p^{i,L}(l))$ then
        $p^{i,L}(l+1) = r^i$
    end
    if $f(r^i(l+1)) > f(p^g(l))$ then
        $p^g(l+1) = r^i$
    end
    $v_n^i(t+\Delta t) = wv_n^i(t) + c_1\phi_1[p_n^{i,L} - x_n^i(t)]\Delta t + c_2\phi_2[p_n^g - x_n^i(t)]\Delta t$
    $r_n^i(t+\Delta t) = r_n^i(t) + \Delta t v_n^i(t)$
end
```

Algorithm 2: Particle swarm algorithm

4 Experimental Results

4.1 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. (Pandey et al., 2012) in their experiments on the Ford dataset (Pandey et al., 2011).
- NMI - normalised mutual information, a metric we had used in our previous work on multi-modal calibration (Taylor and Nieto, 2012).
- The Levinson method (Levinson and Thrun, 2012).
- 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.

For MI, NMI, GOM and the Levinson method, we wrote our own code and optimised each one using the same optimisation process and parameters. The SIFT implementation was taken from code written by Vedaldi and Fulkerson (Vedaldi and Fulkerson, 2010) and the best match was found using a *RANdom SAmple Consensus* (RANSAC) implementation. The SIFT method was only applicable, and thus used, on the first dataset.

4.2 Implementation

The code for testing the methods outlined was written in Matlab, with the exception of the three most computationally expensive sections, these sections were the transformation of the point cloud, interpolation of images and the evaluation of the metrics which were coded in CUDA to reduce computation time[2]. Running on a desktop with an i7-4770 CPU and a GTX760 GPU, the transformation, interpolation and GOM evaluation of a Velodyne scan containing 80,000 points with a camera image requires roughly 4 ms. The process scales linearly with the number of points in the scans and the number of images. However, it is independent of the resolution of the image. To register 20 Velodyne scans with 100 Ladybug images requires approximately 10 minutes for a particle swarm optimisation with 500 particles.

4.3 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.

4.4 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 setup is shown in Figure 10. 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 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 principal distance ≈ 780) and the X, Y and Z coordinates were within 1 metre of correct.

[2] All the code for the implementation as well as documentation is publicly available at http://www.zacharyjeremytaylor.com.

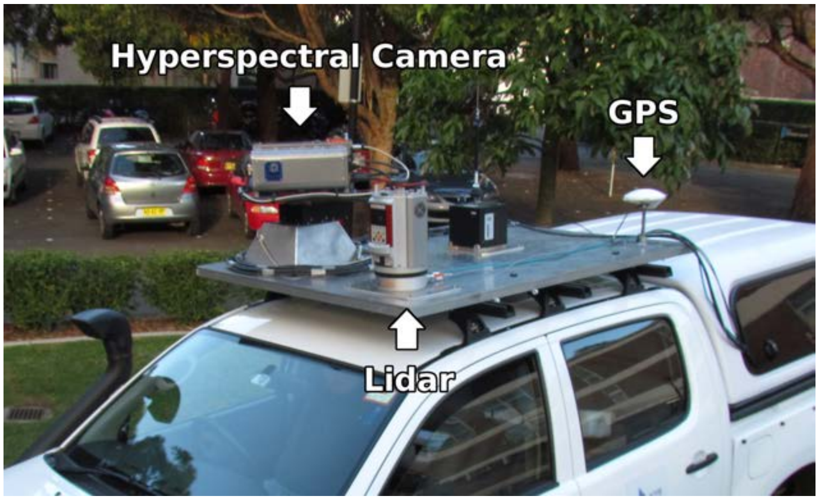


Figure 10: Setup used to collect data around the ACFR for Dataset I

4.4.1 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 example of this is shown in Figure 11. 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 1.

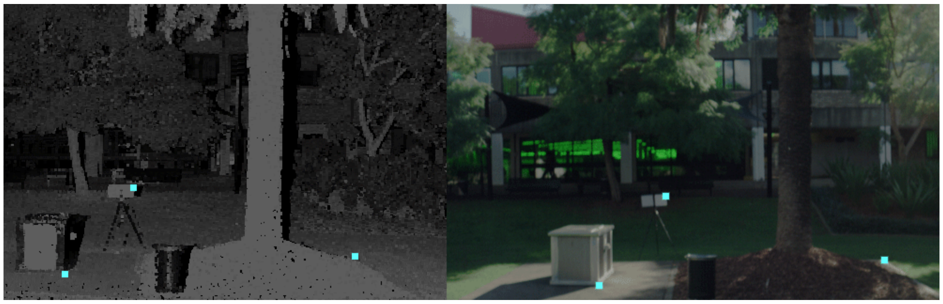
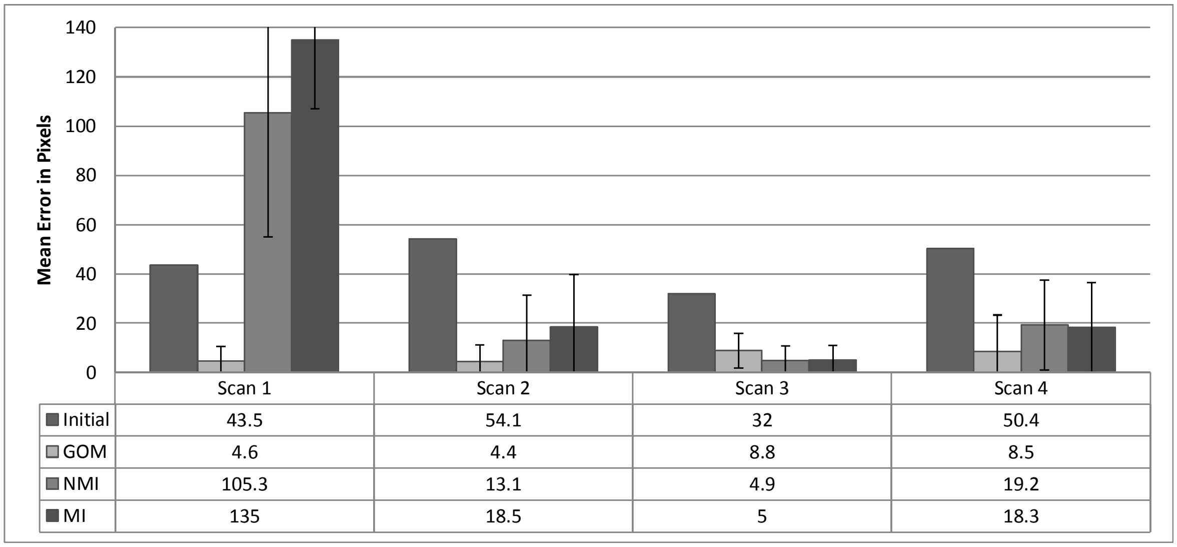


Figure 11: Hand labelled points for a section of ACFR scan 1 aligned by GOM. The left image shows the lidar scan with three of the labeled points highlighted in blue. The right image shows the camera image with the three corresponding points highlighted.

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.



	Scan 1	Scan 2	Scan 3	Scan 4
■ Initial	43.5	54.1	32	50.4
■ GOM	4.6	4.4	8.8	8.5
■ NMI	105.3	13.1	4.9	19.2
■ MI	135	18.5	5	18.3

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

4.5 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. Hyper-spectral cameras capture images from a large number of light wavelengths at the same time. This made them ideal for testing the method's accuracy as the different modality images are initially perfectly aligned as 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 the hyper-spectral camera images, 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 12.

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.

4.5.1 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 2. 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

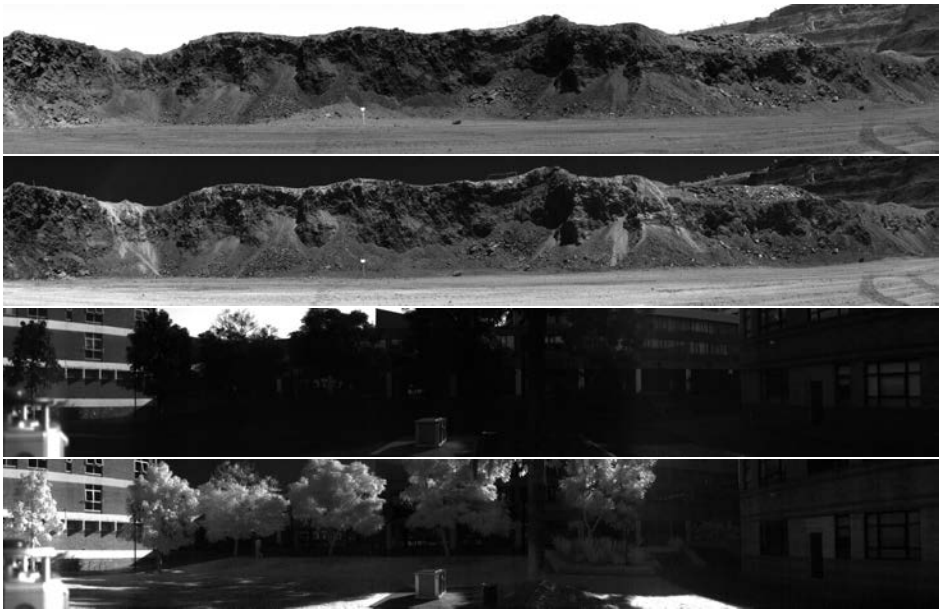
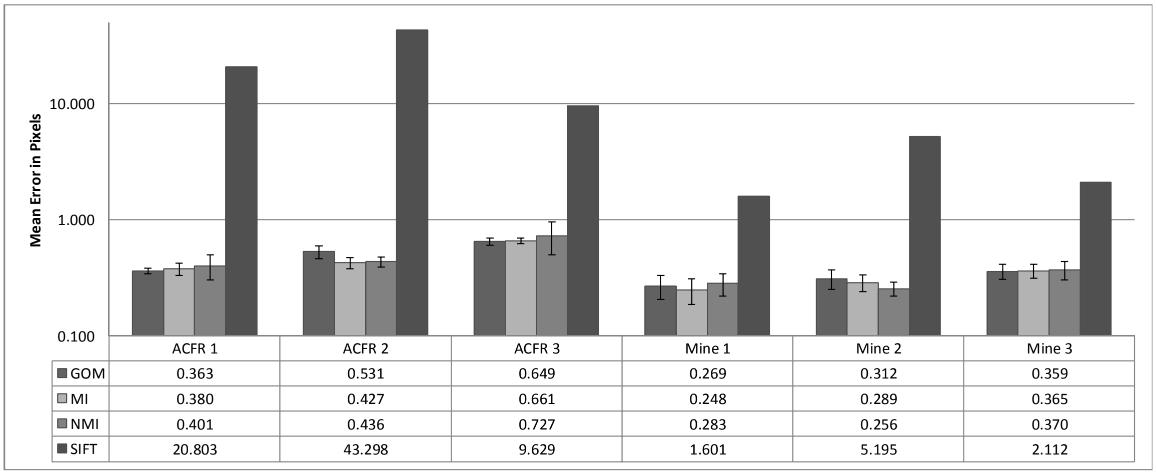


Figure 12: Images captured by hyper-spectral camera. From top to bottom: 420nm mine 1, 950nm mine 1, 420nm ACFR 1, 950nm ACFR 2

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.



	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 2: 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.

4.6 Dataset III

The Ford campus vision and lidar dataset has been published by G. Pandey et al. (Pandey et al., 2011). 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. 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 13. 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.

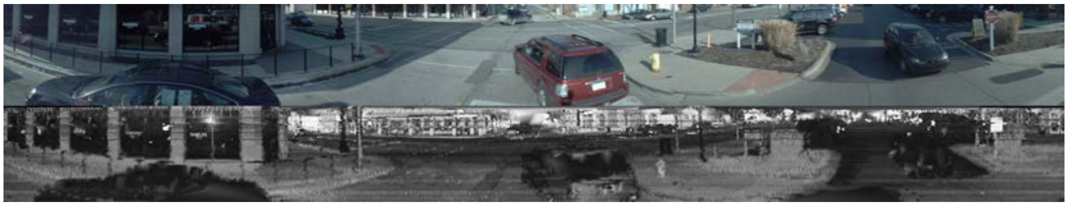


Figure 13: Overlapping region of camera image (top) and lidar scan (bottom) for a typical scan-image pair in the Ford Dataset. The lidar image is coloured by intensity of laser return

4.6.1 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.

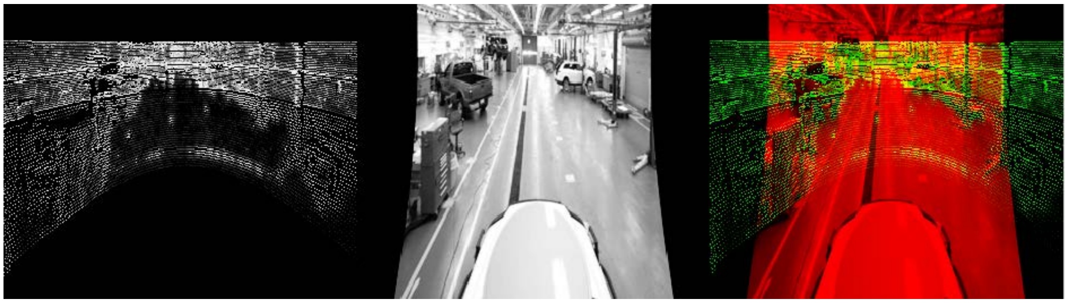


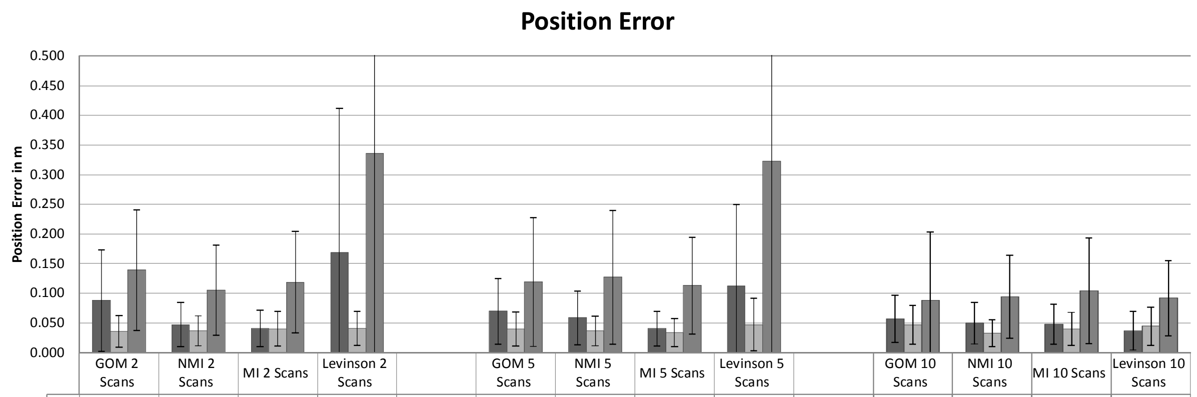
Figure 14: Camera and velodyne scan being registered. Left, the velodyne scan. Centre, the Ladybug's centre camera image. Right the two sensor outputs overlaid.

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 14. 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 (1800 possible combinations) and the average error shown in Table 3.

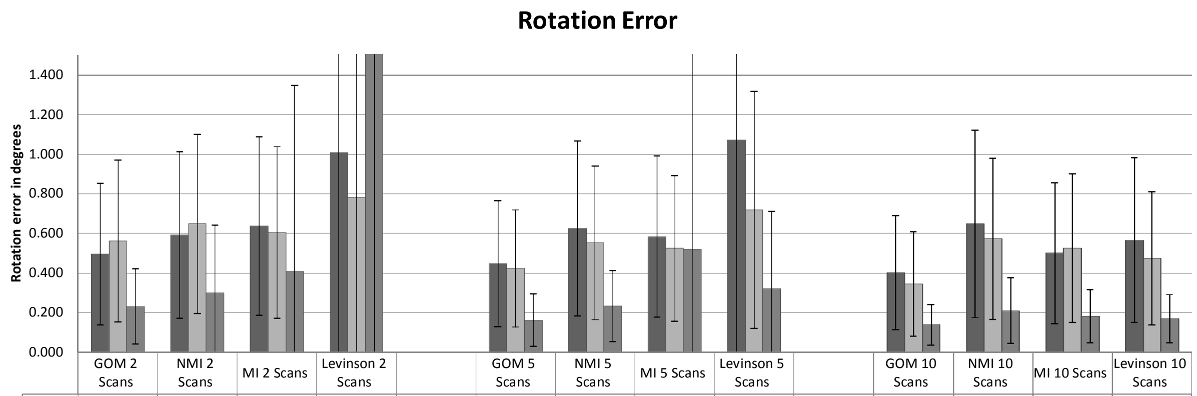
In these tests GOM, NMI and MI gave similar results. GOM tended to give the most accurate rotation estimates while MI gave the most accurate position estimates. For all three of these metrics, scan aggregation slightly improved the accuracy of angles and position. Levinson's presented the largest improvement 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 between the sensors' fields of view.



	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

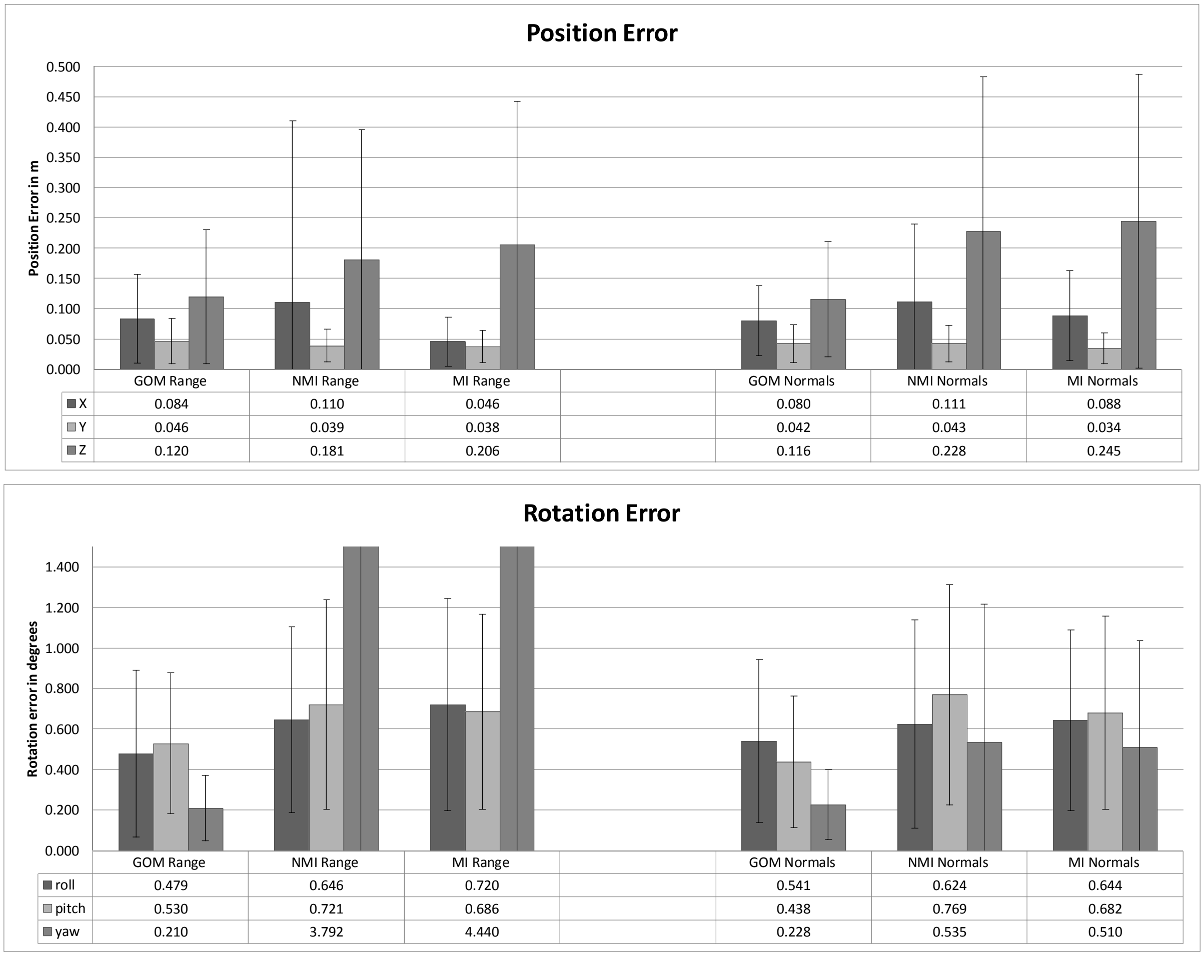


	GOM 2 Scans	NMI 2 Scans	MI 2 Scans	Levinson 2 Scans		GOM 5 Scans	NMI 5 Scans	MI 5 Scans	Levinson 5 Scans		GOM 10 Scans	NMI 10 Scans	MI 10 Scans	Levinson 10 Scans
■ roll	0.496	0.592	0.637	1.007		0.447	0.626	0.584	1.071		0.403	0.649	0.501	0.566
■ pitch	0.562	0.648	0.604	0.783		0.423	0.552	0.524	0.718		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 3: Average error between two aligned Ladybug cameras. All distances are in metres and angles are in degrees.

4.7 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 uncalibrated 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 3.1 must be used. To test their accuracy the Ford dataset was evaluated using these different features, with all other parameters kept the same and 10 scans aggregated for MI, NMI and GOM.



	GOM Range	NMI Range	MI Range		GOM Normals	NMI Normals	MI Normals
■ X	0.084	0.110	0.046		0.080	0.111	0.088
■ Y	0.046	0.039	0.038		0.042	0.043	0.034
■ Z	0.120	0.181	0.206		0.116	0.228	0.245

	GOM Range	NMI Range	MI Range		GOM Normals	NMI Normals	MI Normals
■ roll	0.479	0.646	0.720		0.541	0.624	0.644
■ pitch	0.530	0.721	0.686		0.438	0.769	0.682
■ yaw	0.210	3.792	4.440		0.228	0.535	0.510

Table 4: Average error between two aligned Ladybug cameras for different 3D features. All distances are in metres and angles are in degrees.

The results of these tests are presented in Table 4. For most of the results the metrics gave results that were slightly worse, but otherwise similar to what had been obtained using intensity information. The one exception to this was the yaw angle for MI and NMI when using range as a feature, that showed considerably larger error. While all of these results are worse than those obtained using the return intensity, they are accurate enough to provide a viable option in circumstances where the intensity information is not available.

4.8 Basin of attraction

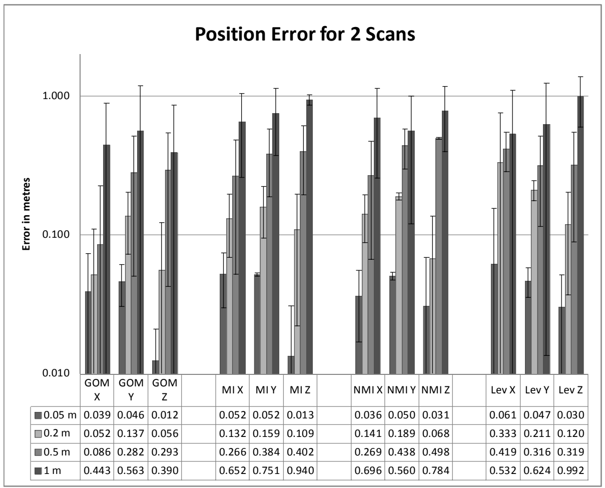
It is preferable to be able to use a local optimisation technique to find the maximum of a metric due to the substantially faster run time. However, as previously stated, the optimiser will only converge to the correct maximum if the initial guess as to the calibration is within its basin of attraction. The size of this region depends on the metric, the number of scans being aggregated and the scene being observed. In an effort to give an indication of the size of the basin for the different metrics, and how scan aggregation affects it, an experiment was performed.

The maximum of a metric was first found by the same method used to test the metrics' accuracy in the Ford dataset. A predetermined error was then added to one of the solutions parameters, and the experiment was re-run using this point as the initial guess. The difference between the initial solution found and the solution found after the error was added were recorded. This was performed for X, Y and Z errors of 0.05, 0.2, 0.5 and 1.0 metres. Roll, pitch and yaw errors of 1, 5, 15 and 30 degrees were also tested. These results were repeated for aggregating 2, 5 and 10 scans. Each experiment was repeated 10 times with random scans.

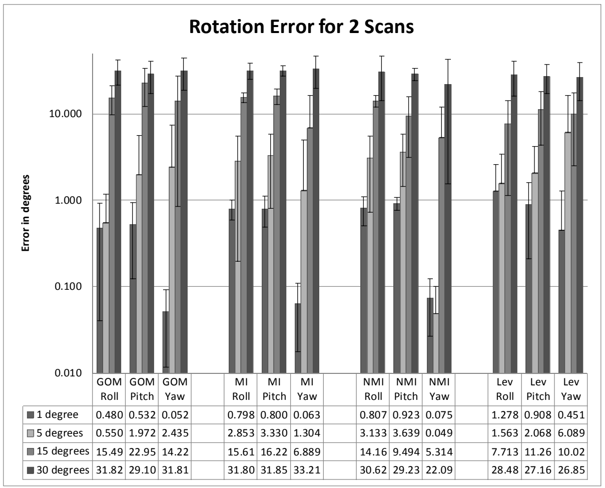
The results of this experiment are presented in Table 5. For the position of the sensor, optimisation significantly reduced the initial error in almost all of the cases. This implies that even when offsets as large as 1 m are present, all four metrics provide an indication of the direction of the correct alignment. However, for low numbers of scans or large offsets while the optimiser reduced the error it still converged to solutions a significant distance from the previously located maximum. When using 2 scans, all X and Y positions showed large error and the Z position required a starting location only 0.05 m from the maximum to converge. The accuracy of all of the solutions increased significantly for 5 scans, and again for 10 scans. This meant, in the 10 scans case both the GOM metric and MI metric provided accurate results for all positions with starting offsets of 0.2 m or less. They also gave accurate results in X for all offsets tested. The Levinson method appeared to give the least consistent results. For example, when 10 scans was used, it generally gave the most accurate X positions. However, occasionally it converged to an incorrect location far from the initial point giving the larger error for the 0.2 m starting offset. Overall, the basin of attraction for GOM and MI appeared to be similar in size, with the Levinson method and NMI's being slightly larger.

In rotation, starting offsets of 15 and 30 degrees often lead to errors larger than before optimisation. This implies that at these angle from the true rotation, the metrics could not provide any strong indication as to the direction of the previously found maximum, and therefore the metrics often converged to an incorrect local maximum in a random direction. While scan aggregation reduced this issue for the 15 degrees offset, it was present in all the 30 degree offset results. The 1 degree offset did not significantly impede any of the metrics from obtaining an accurate result. However using 10 scans, only the GOM solution obtained an accurate result for a 5 degree offset for roll, pitch and yaw. For NMI and MI, significant error was always present in the pitch, while the Levinson method generally had significant pitch and yaw error.

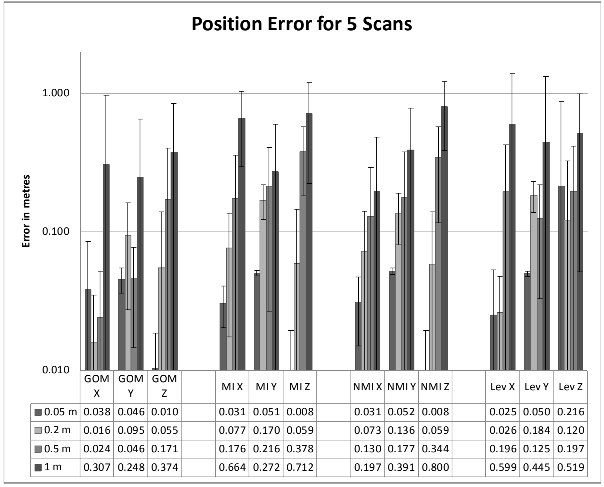
From these results, it can be concluded that all four metrics have a similar basin of attraction, although GOM appeared to handle a slightly wider range of angles than the other methods. In all cases, scan aggregation noticeably improved the results. These results would suggest that for 10 scans on the Ford dataset, a metric must have an initial guess within at least 5 degrees and 0.2 m of the correct solution to obtain reliable calibration results.



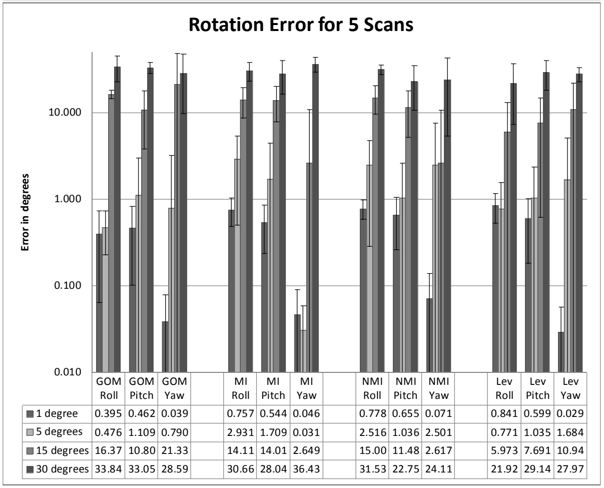
	GOM X	GOM Y	GOM Z		MI X	MI Y	MI Z		NMI X	NMI Y	NMI Z		Lev X	Lev Y	Lev Z
■ 0.05 m	0.039	0.046	0.012		0.052	0.052	0.013		0.036	0.050	0.031		0.061	0.047	0.030
□ 0.2 m	0.052	0.137	0.056		0.132	0.159	0.109		0.141	0.189	0.068		0.333	0.211	0.120
▨ 0.5 m	0.086	0.282	0.293		0.266	0.384	0.402		0.269	0.438	0.498		0.419	0.316	0.319
■ 1 m	0.443	0.563	0.390		0.652	0.751	0.940		0.696	0.560	0.784		0.532	0.624	0.992



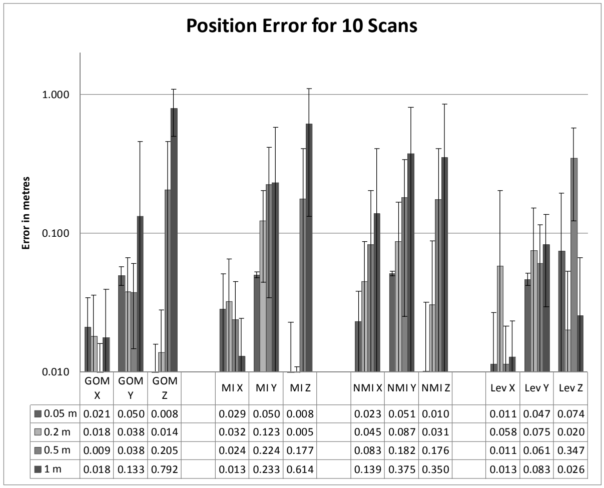
	GOM Roll	GOM Pitch	GOM Yaw		MI Roll	MI Pitch	MI Yaw		NMI Roll	NMI Pitch	NMI Yaw		Lev Roll	Lev Pitch	Lev Yaw
■ 1 degree	0.480	0.532	0.052		0.798	0.800	0.063		0.807	0.923	0.075		1.278	0.908	0.451
□ 5 degrees	0.550	1.972	2.435		2.853	3.330	1.304		3.133	3.639	0.049		1.563	2.068	6.089
▨ 15 degrees	15.49	22.95	14.22		15.61	16.22	6.889		14.16	9.494	5.314		7.713	11.26	10.02
■ 30 degrees	31.82	29.10	31.81		31.80	31.85	33.21		30.62	29.23	22.09		28.48	27.16	26.85



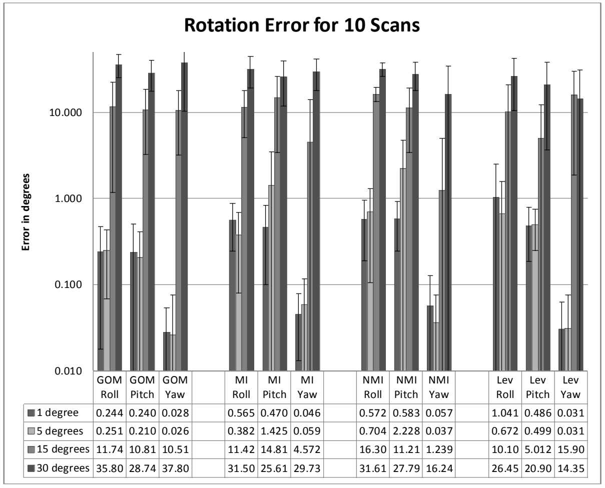
	GOM X	GOM Y	GOM Z		MI X	MI Y	MI Z		NMI X	NMI Y	NMI Z		Lev X	Lev Y	Lev Z
■ 0.05 m	0.038	0.046	0.010		0.031	0.051	0.008		0.031	0.052	0.008		0.025	0.050	0.216
□ 0.2 m	0.016	0.095	0.055		0.077	0.170	0.059		0.073	0.136	0.059		0.026	0.184	0.120
▨ 0.5 m	0.024	0.046	0.171		0.176	0.216	0.378		0.130	0.177	0.344		0.196	0.125	0.197
■ 1 m	0.307	0.248	0.374		0.664	0.272	0.712		0.197	0.391	0.800		0.599	0.445	0.519



	GOM Roll	GOM Pitch	GOM Yaw		MI Roll	MI Pitch	MI Yaw		NMI Roll	NMI Pitch	NMI Yaw		Lev Roll	Lev Pitch	Lev Yaw
■ 1 degree	0.395	0.462	0.039		0.757	0.544	0.046		0.778	0.655	0.071		0.841	0.599	0.029
□ 5 degrees	0.476	1.109	0.790		2.931	1.709	0.031		2.516	1.036	2.501		0.771	1.035	1.684
▨ 15 degrees	16.37	10.80	21.33		14.11	14.01	2.649		15.00	11.48	2.617		5.973	7.691	10.94
■ 30 degrees	33.84	33.05	28.59		30.66	28.04	36.43		31.53	22.75	24.11		21.92	29.14	27.97



	GOM X	GOM Y	GOM Z		MI X	MI Y	MI Z		NMI X	NMI Y	NMI Z		Lev X	Lev Y	Lev Z
■ 0.05 m	0.021	0.050	0.008		0.029	0.050	0.008		0.023	0.051	0.010		0.011	0.047	0.074
□ 0.2 m	0.018	0.038	0.014		0.032	0.123	0.005		0.045	0.087	0.031		0.058	0.075	0.020
▨ 0.5 m	0.009	0.038	0.205		0.024	0.224	0.177		0.083	0.182	0.176		0.011	0.061	0.347
■ 1 m	0.018	0.133	0.792		0.013	0.233	0.614		0.139	0.375	0.350		0.013	0.083	0.026



	GOM Roll	GOM Pitch	GOM Yaw		MI Roll	MI Pitch	MI Yaw		NMI Roll	NMI Pitch	NMI Yaw		Lev Roll	Lev Pitch	Lev Yaw
■ 1 degree	0.244	0.240	0.028		0.565	0.470	0.046		0.572	0.583	0.057		1.041	0.486	0.031
□ 5 degrees	0.251	0.210	0.026		0.382	1.425	0.059		0.704	2.228	0.037		0.672	0.499	0.031
▨ 15 degrees	11.74	10.81	10.51		11.42	14.81	4.572		16.30	11.21	1.239		10.10	5.012	15.90
■ 30 degrees	35.80	28.74	37.80		31.50	25.61	29.73		31.61	27.79	16.24		26.45	20.90	14.35

Table 5: Average error in optimisation for different levels of initial offset. All distances in metres and angles in degrees. Note the chart axis uses a log scale

5 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 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 have also explored three different features (return intensity, range and normals) that could be used in combination with GOM or the mutual information techniques tested to register the outputs. Finally, we examined the level of accuracy required for an initial guess for a system's calibration to be optimised to the correct solution.

From the experiments performed and using all of the different methods we have found that there are situations where each of the methods' strengths make it the most appropriate to use. In all tests conducted, NMI was generally outperformed by other methods, giving little reason to use it for the type of problems examined. For the calibration or recalibration of a Velodyne-camera setup or similar system, all four metrics give similar, accurate results. However, if a large number of scans are available, Levinson's method will present accurate results without requiring any form of intensity information. It also has the advantage of requiring less memory and being more quickly computed due to only using points near large depth discontinuities. This method, however, requires large numbers of scans to obtain accurate results. If these are not available, both MI and GOM are better alternatives, MI being faster to compute. Finally, for the registration of one-off high resolution scans with images, GOM was the only metric that consistently provided accurate results.

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

Audet, C. and Jr, J. D. (2002). Analysis of generalized pattern searches. *SIAM Journal on Optimization*, 13(3):889–903.

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

Bouguet, J. (2004). Camera calibration toolbox for matlab.

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.

Chen, S. (2009). Another Particle Swarm Toolbox.

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

Goldberg, D. and Holland, J. (1988). Genetic algorithms and machine learning. *Machine learning*, pages 95–99.

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.

Kumar, R. and Ilie, A. (2008). Simple calibration of non-overlapping cameras with a mirror. *CVPR*.

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.

Lébraly, P. and Royer, E. (2011). Fast calibration of embedded non-overlapping cameras. *ICRA*.

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.

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.

Mikki, S. M. and Kishk, A. a. (2008). *Particle Swarm Optimization: A Physics-Based Approach*, volume 3.

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)*.

Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*.

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.

Penney, G. P., Weese, J., Little, J. a., Desmedt, P., Hill, D. L., and Hawkes, D. J. (1998). A comparison of similarity measures for use in 2-D-3-D medical image registration. *IEEE transactions on medical imaging*, 17(4):586–95.

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.

Roche, A., Malandain, G., Pennec, X., and Ayache, N. (1998). The correlation ratio as a new similarity measure for multimodal image registration. *MICCAI*.

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.

Ugray, Z. and Lasdon, L. (2007). Scatter search and local NLP solvers: A multistart framework for global optimization. *INFORMS Journal*.

Unnikrishnan, R. and Hebert, M. (2005). Fast extrinsic calibration of a laser rangefinder to a camera.

Vedaldi, A. and Fulkerson, B. (2010). VLFeat: An open and portable library of computer vision algorithms. *ACM International Conference on Multimedia*.

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.

Wang, R., Ferrie, F. P., and Macfarlane, J. (2012). Automatic registration of mobile LiDAR and spherical panoramas. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40.

Zana, F. and Klein, J. C. (1999). A multimodal registration algorithm of eye fundus images using vessels detection and Hough transform. *IEEE transactions on medical imaging*, 18(5):419–28.

Zhang, Q. and Pless, R. (2004). Extrinsic calibration of a camera and laser range finder (improves camera calibration). *Intelligent Robots and Systems*, pages 2301–2306.