Towards Multimodal and Condition-Invariant Vision-based Registration for Robot Positioning on Changing Surfaces

James Sergeant[1], Gary Doran[2], David R. Thompson[2], Abigail Allwood[2], Chris Lehnert[1], Ben Upcroft[1] and Michael Milford[1]

[1] ARC Centre of Excellence for Robotic Vision
School of Electrical Engineering and Computer Science
Queensland University of Technology, Brisbane, Australia

[2] The Jet Propulsion Laboratory
California Institute of Technology
Pasadena, United States.

Abstract

Autonomously positioning a robot arm toolpoint, mobile base or UAV relative to a surface of interest is a fundamental capability in many applications such as infrastructure monitoring and planetary rock inspection. Robots performing this task in uncontrolled real world environments face many challenges such as adverse weather conditions, changing illumination or even the need to perform positioning using multiple sensing modalities. In this paper, we address this key robotics challenge in the context of the Mars2020 rover mission. Scientists require the ability to manually identify a target on a rocky outcrop using a centrally mounted RGB sensor and command the rover to autonomously position a robot arm toolpoint at that target on the following day, relying only on the x-ray sensor mounted on the arm toolpoint itself. We use an adapted sequence-based technique for multi-modal image registration that builds on recent appearance-invariant robotic place recognition algorithms. We introduce a new large, custom dataset of rock samples with multimodal sensing observations, and compare the performance of the proposed technique to a convolutional neural network (CNN)-based approach as well as a traditional Speeded-Up Robust Features (SURF)-based approach. Finally, we demonstrate the entire system performing vision-based positioning of a robot arm tool point on actual rock samples with significant illumination change.

1 Introduction

If autonomous robots are ever to be a permanent presence in everyday life, they must be able to navigate, sense and interpret the environment around them. One of the key challenges in achieving this aim is the fact that the world constantly changes, due to day-night cycles, weather cycles, seasonal change, and dynamic objects in the environment. In recent years, this problem has received particular attention in the navigation domain, where a range of approaches have been developed to deal with environmental change, including using learning techniques [Biederman, 1988; Johns & Yang, 2013; Lowry *et al.*, 2014; Neubert *et al.*, 2013], condition-invariant 'features' [Milford *et al.*, 2014b; Ranganathan *et al.*, 2013], an illumination invariant colourspace [Corke *et al.*, 2013; Maddern *et al.*, 2014], data association flow networks [Naseer *et al.*, 2014], naive Bayes nearest neighbour scene descriptors [Kanji, 2015] or integrating information over temporal sequences of images [Arroyo *et al.*, 2015; Milford, 2013; Milford & Wyeth, 2012; Pepperell *et al.*, 2014]. However, there is far less research devoted towards the parallel problem of detecting, recognizing and interacting with surfaces of interest in the environment under similar extreme degrees of perceptual change.

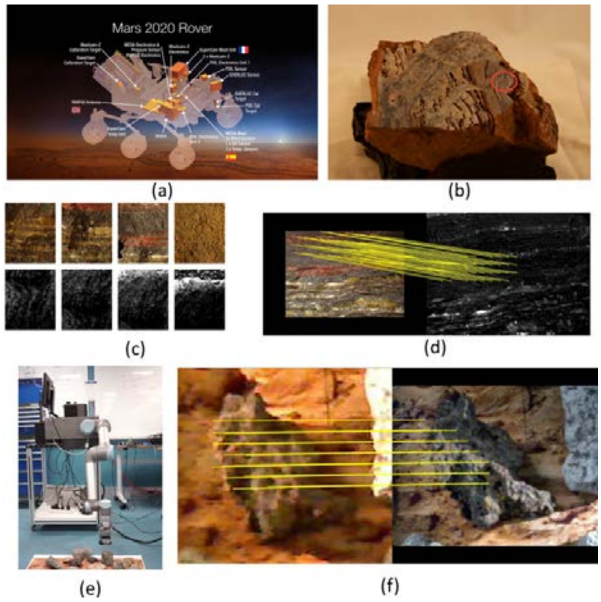

Figure 1: We present a novel sequence-based method for appearance-invariant image registration and demonstrate it outperforming CNN- and SURF-based techniques on a novel multi-modal sensor rock dataset, as well as in autonomous experiments in a Mars-like environment.

In this paper, we present new research that addresses these problems in the context of the Mars2020 rover mission. Mars2020 is a Mars rover mission planned for launch in 2020, with the primary goal to investigate the possibility of past and extant life on Mars. These investigations generally involve interrogating rock outcrops with arm-mounted sensors. To efficiently deploy these sensors, humans will direct the rover toward target points that they manually identify in high resolution RGB outcrop images; this typically only occurs once daily. These context images will initially be obtained from a centrally located high-resolution, stereo imaging camera mast (Mastcam-Z) [Novak *et al.*, 2015]. After receiving the command, the rover should automatically servo arm-mounted sensors such as the Planetary Instrument for X-Ray Lithochemistry (PIXL) [Allwood *et al.*, 2015], an x-ray fluorescence spectrometer, to precisely measure the selected target point. This task is extremely challenging for several reasons. First, the relative pose of the two sensors cannot be guaranteed to be known accurately. Second, the actual visual servoing may occur much later at a completely different time of "day" with very different lighting conditions. Thus, the two sensors create very different images (the multimodal registration challenge). Thirdly, the ideal solution will be able to perform this task directly based on imagery of the rock sample, rather than by using artificial visual markers or beacons. Finally, all this must be achieved under the restricted computational environment of radiation hardened, spaceflight qualified processors, the performance of which lags far behind terrestrial CPUs.

In beginning address these challenges, we are continuing to develop a custom sequence-based image registration technique that leverages the appearance-invariant properties of sequences of locally-normalized patches within the query/reference image pair to be registered. To evaluate performance, we have gathered a large new multi-modal rock database, with sensors that mimic the properties of the PIXL and other sensors proposed for the Mars2020 mission. Finally, we present initial mock up trials that test the ability of the system to correctly position a robot toolpoint relative to real rock samples under significant lighting changes. Although our initial implementation is not yet computationally optimized, we provide computational calculations that demonstrate the approach is feasible on constrained computational equipment.

The paper proceeds as follows. In Section II we provide a brief literature review on methods for image registration, background information on invariant visual recognition and visual servoing techniques, and describe the technical details of the Mars2020 application scenario. Section III presents an overview of the approach taken, with Section IV describing the experimental setup. Section V presents the results of the database and real robot experiments, with discussion and conclusion in Section VI.

2 Background

2.1 Mars2020 Mission

Mars2020 is a planned rover mission with the primary aim of investigating the habitability of ancient Martian environments. Mission studies will include investigation of the geological processes with respect to its past and current habitability. One of the primary mechanisms by which this will be achieved is collection and analysis of rock cores and soil samples. Like many planetary missions, the acquisition will be semi-autonomous: operators on Earth will analyse high resolution imagery from the rover's onboard cameras and identify regions of interest. The rover will autonomously servo a sensor-equipped arm to the target location and perform tasks such as sample collection.

Many of the problems faced in similar scenarios on Earth are exacerbated on Mars: the relative pose of the two sensors is unknown and sensor calibration may drift. Due to operational requirements, actual servoing of the arm may occur at a later time when lighting conditions are drastically different. Finally, a target area identified in the image produced by one type of sensor (such as an RGB camera) must ideally be recognizable using sensory input from another completely different sensing modality (such as an x-ray sensor) at different resolutions and relative poses.

2.2 Visual Servoing Under Challenging Conditions

Visual servoing has been used for decades [Corke, 1996; Malis & Rives, 2003] and has been exploited in a wide range of robotic applications such as industrial manufacturing [Kosmopoulos, 2011], mobile robot navigation [Bonin-Font *et al.*, 2008], medical robotics [Azizian *et al.*, 2014], control of UAV platforms [Sa *et al.*, 2015], grasping in cluttered environments [Leeper *et al.*, 2014] and other manipulation tasks [Kragic & Christensen, 2002]. This body of work has mostly been conducted under idealised or controlled lighting conditions. In contrast, there has been relatively little progress made in visual servoing for challenging and outdoor environments. Using visual information for robot servoing is challenging as it relies on the accuracy of a camera's calibration, and because a model of the object's geometry or image features can be a highly non-linear function of the camera pose. Furthermore, real world problems of lighting, stability and computational performance make visual servoing difficult.

Some examples where visual servoing has been applied to perceptually challenging environments include agriculture, such as fruit harvesting [Mehta & Burks, 2014] or weed destruction [Michaels *et al.*, 2015]. Another challenge is robustness to camera calibration [Corke, 2011]. For instance, computing pose errors for visual servoing within a robot's operational space instead of the image space is sensitive to sensor calibration errors and would not be appropriate for application on the Mars2020 mission [Corke, 2011].

Multi-sensor visual servoing incorporating sensor fusion has been demonstrated previously however relies on intersensor calibration, artificial markers for tracking of the manipulator-mounted camera and was not demonstrated with different types of camera sensors or under challenging visual conditions [Kermorgant & Chaumette, 2011].

2.3 Invariant Visual Place Recognition Techniques

In the place recognition domain, a range of new algorithms have been developed that address the problem of appearance-invariant place recognition. These approaches span learning techniques [Biederman, 1988; Johns & Yang, 2013; Lowry *et al.*, 2014; Neubert *et al.*, 2013], condition-invariant patch 'features' or integrating information over temporal sequences of images [Milford, 2013; Milford & Wyeth, 2012; Pepperell *et al.*, 2014]. Sequence-based techniques in particular have demonstrated state of the art appearance-invariance, if the relative pose is known approximately. These characteristics suit the Mars2020 rover mission scenario, where approximate relative pose of the various sensors is likely to be known and great environmental appearance-change is expected. Consequently, the methods presented here build on the foundation of sequence-based navigation techniques, but leverage their appearance-invariance to perform matching *within* an image, rather than across multiple images in a video stream.

3 Approach

Two methods for generating feature correspondences, SeqSLAM and a convolutional neural network (CNN)-based approach, were investigated to determine their suitability for aligning images within a multimodal image dataset of rock samples provided by the PIXL sensor investigation team at NASA JPL. A third standard feature detector, Speeded-Up Robust Features (SURF), was also investigated for its suitability for the dataset and as a baseline performance measure. The dataset contains RGB and x-ray sensor images of 22 different rock samples with differences in direction of illumination, camera position and focal length (scale).

3.1 Intra-Image Best Match Image Registration

The vanilla SeqSLAM algorithm performs place recognition by integrating single frame place match hypotheses over many frames in order to produce an overall sequence matching hypothesis [Milford & Wyeth, 2012]. In this work, in order to generate feature correspondences between two images, we have further adapted an intra-image implementation of SeqSLAM, first introduced in [Milford *et al.*, 2014a] which we apply to virtual navigation sequences within the two images (Figure 2). Here we apply the method across different modes of image data (RGB and infra-red) and use it in robot experiments under extreme changes in lighting conditions.

Raw images are first patch-normalized and resolution reduced. Each image is then divided up into a grid, and virtual navigation sequences with *q* stops along a trajectory of length *z* in the query image are correlated across the entire reference image to produce a sequence of difference matrices. These difference matrices are aligned based on the search trajectory and then cropped to a sub-region difference matrix. The minimum difference score within this matrix is used as the matching location to generate the feature pair.

3.2 Image Pre-Processing

Initially, the images were pre-scaled by scaling factors *SF* based on available information about the sensor focal extension *f* (or equivalent) and an estimate of the distance *d* between the sensor and the sample surface.

$$d_{\max} = \max\left(d_{query}, d_{reference}\right) \tag{1}$$

$$f_{\min} = \min\left(f_{query}, f_{reference}\right) \tag{2}$$

$$SF_{query} = \frac{d_{query}}{d_{\max}} \frac{f_{\min}}{f_{query}} \tag{3}$$

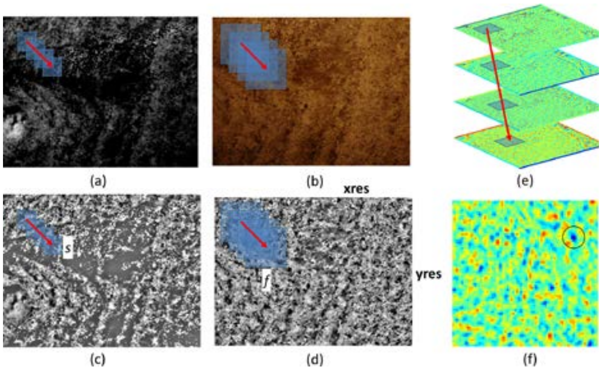

Figure 2: Intra-image SeqSLAM. After the (a-b) raw images are patch-normalized, q patches evenly spaced along a matching trajectory of length z in the (c) new image are correlated across the (d) entire reference image to produce a (e) sequence of difference matrices. These difference matrices are aligned based on the search trajectory and then cropped to a (f) sub-region difference matrix. The minimum difference score within this matrix is used as the matching location.

$$SF_{reference} = \frac{d_{reference}}{d_{\max}} \frac{f_{\min}}{f_{reference}} \tag{4}$$

Images were then resampled at reduced resolution and patch normalization applied to reduce the effects of local variations in illumination and appearance changes across multiple sensory modes (Figure 2a-d). The patch normalized pixel intensities, I', are given by $I'_{xy} = \left(I_{xy} - \mu_{xy}\right)/\sigma_{xy}$, where μ_{xy} and σ_{xy} are the mean and standard deviation of pixel values in a patch of size P_{size} surrounding (x, y). The code and datasets will be made freely available at the following link: https://wiki.qut.edu.au/display/cyphy/Datasets.

Algorithm 1 provides an algorithmic description of the full intra-image SeqSLAM image registration algorithm process:

Algorithm 1: SeqSLAM Algorithm

```
Convert query and reference images to greyscale and pre-scale
Resize larger image to maximum dimension $xres$ or $yres$, as appropriate
Scale the smaller image by the same ratio
Perform local patch normalisation with patch size *nps* and minimum
    standard deviation *minstd*
For each trajectory 1 to *n*:
        For point 1 to *q* in trajectory:
                Extract region of size *s* about point in the query
                image
                Perform sum of absolute difference between region
                    and reference image
        Align and crop the resulting *q* difference maps with respect
            to the search trajectory
        Sum across all difference maps
        The point of minimum difference is taken as the matching
            point to the first point on the trajectory
Use MSAC to estimate a suitable similarity transform
```

3.3 CNN-based Image Registration

The second approach utilises off the shelf unsupervised machine learning methods trained on existing large image databases (future work discusses training specifically for the domain). We assume that significant features in a sensor image will return a large response when summing across the feature maps generated by the first convolutional layer. Effectively, we use the learned low-level features of AlexNet as an interest point detector. Algorithm 2 describes the CNN-based image registration method process.

Algorithm 2: CNN-based Image Registration Algorithm

```
Pre-training:
Use an unsupervised training method to learn suitable convolutional
    filter parameters, initially using large image databases but
    eventually on WATSON (RGB) and PIXL (greyscale) sensor
    datasets
In use:
Pre-scale query and reference images
Separately, for the query and reference images:
        Pass image through the first convolutional and rectified linear
            unit layers of trained network using RGB or greyscale
            convolutional layer parameters, as required
        Sum the resulting $n_{filters}$ feature maps
            $$M_{summed}(u,v) = \sum_{i}^{n_{filters}} M_{n_{filters}}(u,v)$$
Separately for $M_{summed,query}$ and $M_{summed,reference}$, for 1 to $n_{peaks}$:
        Select point of maximum response in $M_{summed}$
        Extract region of size $2rw_{image}$ from around the point in
            $M_{summed}$
        Ignore other peaks in the region extracted
For the $n_{peaks}$ regions extracted from $M_{summed,query}$ and
    $M_{summed,reference}$, perform feature matching using sum of squared
    differences
Use MSAC to estimate a suitable similarity transform
```

4 Experimental Setup

This section describes the experimental setup, dataset acquisition and pre-processing and key parameter values.

4.1 Rock Dataset

The Mars2020 rover will have a large array of sensors. Of particular interest to this work are the Mastcam-Z high resolution, stereo imaging camera, the WATSON microscopic context imager, the PIXL instrument and the PIXL microscopic context camera (PIXL-MCC) imager along with their test analogues [Novak *et al.*, 2015]. The dataset (Figure 3) developed by the NASA JPL PIXL sensor investigation team contains a total of 198 RGB images (WATSON [Novak *et al.*, 2015]) and 418 x-ray sensor images (PIXL, 16-bit greyscale) for 22 rock samples taken with a variety of differences in the direction of illumination (both), sensor position (PIXL) and focal extension (WATSON, scale / field of view) when acquired. Because the actual mission sensors are in development, we used analogous imaging sensors to precisely match the optical and resolution properties of the design specification. Given the translation of the camera in most instances excludes the possibility of a match, the experimental test set was constructed such that the rules in Table 1 were observed between two images.

A total of 5368 test cases were generated in 14 different categories as shown in Table 2. The categories were determined by the mode of each of the images and whether there is a difference in translation, lighting, scale or mode between the two images. Given that each of the methods uses information about one of the images to find feature matches within the other, the multimodal categories were intentionally kept separate.

4.2 Match Criteria

Given that each of the methods are provided with approximate relative scale, the expected scale component of the geometric transform estimated should be approximately 1. Based on a zero-rotation assumption, a second match criteria required there be less than 10° rotation when applying the transform. A final criterion considered the magnitude and direction of known translation in the x and y directions. Based on the difference in translation between the two images, this criterion reduces false-positive matches that were not detected using the previous criteria.

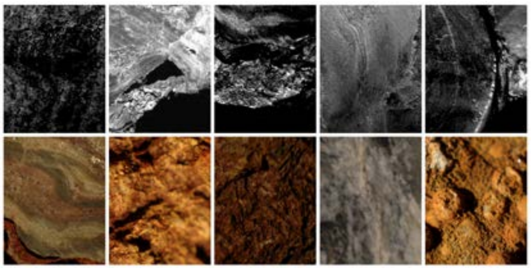

Figure 4: Multi-modal rock sample images from the PIXL (greyscale) and WATSON (RGB images) test dataset.

Table 1: Test Set Construction Rules

Rule	Rationale
Only match images from the same context	In practice, two images from different samples will not be matched.
Only test images with translation equal to or less than 0.5 "units" in any direction	For distances greater than 0.5 "units", the overlapping region between the two images is quite small and unlikely allows for a reasonable match. In multimodal cases with a large scale (focal extension) difference, this is even more significant.
No matching required between the same image	Trivial case.

Table 2: Test Categories

Category	Mode		Differences				Total
	Image 1	Image 2	Translation	Lighting	Scale	Multimodal	
1	WATSON	WATSON	False	False	True	False	**396**
2	WATSON	WATSON	False	True	False	False	**396**
3	WATSON	WATSON	False	True	True	False	**792**
4	WATSON	PIXL	False	False	True	True	**198**
5	WATSON	PIXL	False	True	True	True	**396**
6	WATSON	PIXL	True	False	True	True	**264**
7	WATSON	PIXL	True	True	True	True	**528**
8	PIXL	WATSON	False	False	True	True	**198**
9	PIXL	WATSON	False	True	True	True	**396**
10	PIXL	WATSON	True	False	True	True	**264**
11	PIXL	WATSON	True	True	True	True	**528**
12	PIXL	PIXL	False	True	False	False	**132**
13	PIXL	PIXL	True	False	False	False	**704**
14	PIXL	PIXL	True	True	False	False	**176**
						Total	**5368**

4.3 SURF Comparison

The MATLAB implementation (with its default parameters) of the SURF detector [Bay *et al.*, 2006] was utilised as a baseline method in image registration of the PIXL-WATSON dataset. The SURF detector finds blob-like features through the application of the Hessian matrix at a variety of scale levels [Bay *et al.*, 2006]. To improve performance, the Hessian is comprised of box filter approximations of second order Gaussian derivatives [Bay *et al.*, 2006]. The determinant of the Hessian allows for scale selection [Bay *et al.*, 2006]. Prior to application, the images were converted to greyscale and pre-scaled. The detector was then applied to both images separately and matching performed on the extracted features for use in estimation of an appropriate transform.

4.4 CNN Comparison

In other recognition domains, such as visual place recognition, pre-learnt features have been found to perform at better than state-of-the-art levels. To investigate their feasibility we used the low-level convolutional filters learned by AlexNet [Krizhevsky *et al.*, 2012], trained on the ImageNet dataset and provided in the distribution of Caffe. Given that low-level features across many CNNs are typically demonstrate similar characteristics, AlexNet was used for ease of implementation.

4.5 Robot Experiments

As a proof of concept, robot experiments were undertaken to demonstrate the method in application of tool point positioning under changing illumination conditions using a vision sensor. Here we used a Universal Robots UR5 robot arm, a lightweight robot arm with a working radius of 850 mm. The arm was equipped with a Logitech webcam mounted on the toolpoint, with the entire robot located facing a rock field on top of a Mars diorama. Higher fidelity testing environments are currently under development; this current setup enables the testing of the illumination invariant properties of the approach and a trial of the overall end to end system operation.

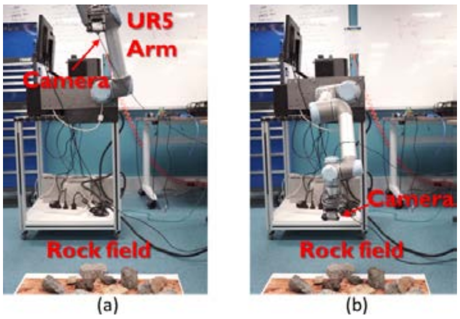

Figure 5: UR5 robot arm in its test configurations; (a) mimicking the centrally located camera on the rover and (b) simulating the location of the PIXL sensor at the front of the rover.

Experiments consisted of the robot arm first being positioned in a "stand-off" location with the camera approximately 1.5 m distant from the rock field, to mimic the centrally located camera on the rover (Figure 5a). A snapshot was taken and an operator manually tagged a region of interest within that image. The toolpoint was then randomly relocated to a position in the general vicinity of the rock field (Figure 5b), to simulate the location of the PIXL sensor at the front of the rover. For the experiment and for both sensor positions, a distance estimate is made by adding a random value to the known distance; in a real system, this distance could be estimated by a human operator or a depth sensor. At this point, image registration was performed using the SeqSLAM technique to obtain an approximate relative pose estimate, and the tool point was autonomously positioned at the approximate target location. The scale returned from the registration method enabled a better approximation of the distance between the object and the sensors, allowing for a reasonable estimate of an appropriate extension of the arm towards the sensor. Additionally, the translation components of the image alignment transform and the sensor's field of view enabled the estimation of the required translation of the camera towards the region of interest.

Experimental Process 1: Robot Experiments

```
Initialise Experiment:
        Random Reference Position:
                $x_{reference} = N(0,0.05)$
                $y_{reference} = N(-0.6,0.05)$
                $z_{reference} = N(1.5,0.05)$
        Random Sensor Position:
                $x_{sensor} = N(0,0.1)$
                $y_{sensor} = N(-0.6,0.05)$
                $z_{sensor} = N(0.7,0.15)$
        Estimated Distances:
                Experimental error $e = N(0,0.1)$
                Estimated $d_{Ref-Sample} = z_{reference} + e$
                Estimated $d_{Sensor-Sample} = z_{sensor} + e$
Process:
Acquire reference image, operator selects region of interest
Estimate of reference to sample surface $d_{Ref-Sample}$
Calculate $d_{ROI-Sample} = \frac{d_{Ref-Sample} \cdot w_{ROI}}{w_{Ref}}$, where $w$ is width in pixels
Using known poses of sensors, extrapolate sensor to sample surface
    $d_{Sensor-Sample}$
Calculate relative scales $SF_{Sensor} = SF_{query}$ and $SF_{ROI} = SF_{reference}$
Perform image registration (as per Algorithm I) using the calculated
    image ratios for pre-scaling and returning $T$, the image alignment
    transform
Extract scale, rotation and translations in x and y directions the image
    alignment transform, where
                $scale = \sqrt{a^2 + b^2}$
                $rotation = atan2(-b, a)$
Estimate new sensor pose:
        Update ROI-sample surface estimate:
                $d_{ROI-Surface} := d_{ROI-Surface} \cdot scale$
        $dZ = (d_{Sensor-Surface} - d_{ROI-Surface})$
        Using sensor properties
                $A = \frac{2 \cdot d_{ROI-Surface} \cdot tan\left(\frac{FOV_{sensor}}{2}\right)}{w_{image}}$
        $dX = A \cdot t_X$
        $dY = A \cdot t_Y$
        $d\theta = rotation$
        $[X, Y, Z, \theta]_{new} = [X, Y, Z, \theta]_{current} + [dX, dY, dZ, d\theta]$
        $d_{Sensor-Surface} := d_{ROI-Surface}$
        where
                $dZ$ : change in extension normal to sample
                    surface
                $dX, dY$: translational components with respect
                    to sample surface
                $FOV_{sensor}$ is sensor field of view in radians
                $A$ is pixel to real-world distance conversion
                    factor
                $t_X$, $t_Y$ and $w_{image}$in pixels
Move to estimated pose
For ideal lighting experiments:
        After move to first estimated pose, use PID control to
            minimise the estimated scale, rotation and translation
            components of the image registration transform returned
            by the SURF method
For lighting difference experiments:
        Repeatedly estimate new sensor pose (steps 7 and 8) using
            relative sensor image scales of 1:1
```

With the sensor moved to the initial estimate target pose, the toolpoint was visually positioned using a PID controller to minimise the scale difference, rotation and translation between the region of interest image and the sensor's current view. The system stopped after achieving its halting criterion of less than 1% scale difference, 1° rotation and 5 pixels translation. Additionally, the positioning stopped if the image registration was deemed unsuccessful. For the initial experiments, the SURF image registration method was utilised. For the changing lighting experiments, the SURF approach failed; instead SeqSLAM was used for both stages of image registration along with open-loop control and looser stopping criteria to less than 3% scale difference, 3° rotation and 15 pixels' translation.

Ground truth was unavailable for the equivalent robot pose of a selected region of interest; assessment of the final toolpoint positioning was made by a human operator based on the alignment of the final viewpoint with the region of interest with at least 90% overlap by visual inspection. Additionally, the average pixel distance between 4 points in the region of interest and the final view point is reported.

4.6 Robot Experiment Algorithm

Experimental Process 1 outlines the method for initialising the UR5 sensor poses, making the initial distance estimate and the step-by-step process for acquiring sensor images, performing image registration and positioning towards the region of interest.

4.7 Parameter Values

Table 3 provides the values and descriptions for the experimental parameters used in each of the image registration methods in this paper and Table 4-6 detail the camera's specifications, the PID coefficients and the stopping criteria used in the robot experiments. The parameters were selected based on those provided by [Milford *et al.*, 2014a]; however it is expected that optimisation for computation and performance may be achieved through intelligent parameter search.

Table 3: Experimental Image Registration Parameters

SeqSLAM Image Registration		
Parameter	**Value**	**Description**
P_{size}	11x11	Dimensions of patch for local patch normalisation
σ_{min}	0.1	Minimum standard deviation used for local patch normalisation
b	0.1	Ignore cases within b × image dimension of image border
s	41×41	Width of patch comparison area in pixels
xres,yres	400	Length of maximum dimension of patch normalized image in pixels
q	6	Number of patches used in the intra-frame sequence search
z	100	Horizontal and vertical length components of diagonal trajectory
step	20	Step size (in *u,v* directions) between trajectory starting points
CNN-based Image Registration		
Parameter	**Value**	**Description**
$n_{filters}$	96	Number of convolutional filters
n_{peaks}	1000	Number of peaks to extract from the summed feature maps
r	0.02	Patch radius (as a fraction of the maximum image dimension) for extraction about response peaks
SURF Image Registration (MATLAB implementation)		
Parameter	**Value**	**Description**
MetricThreshold	1000	Number of strongest features selected
NumOctaves	3	Number of octaves
NumScaleLevels	4	Number of scale levels to compute per octave

Table 4: Camera Specifications

Parameter	**Value**	**Description**
h_{Ref} or h_{Sensor}	720	Height of reference and sensor images in pixels
w_{Ref} or w_{Sensor}	960	Width of reference and sensor images in pixels
FOV_{sensor}	54	Horizontal field of view in degrees

Table 5: PID Coefficients

Parameter	**Scale**	**Rotation**	**Translation (x & y)**
k_p	0.05	0.5	0.00005
k_i	0	0	0
k_d	0	0	0.00001

Table 6: Stopping Criteria

Experiment	**Scale**	**Rotation**	**Translation**
Ideal Lighting	± 1%	± 1°	± 5px
Lighting Difference	± 3%	± 3°	± 15px

5 Results

In this section, the results of each of the image registration methods are presented along with examples of feature / region matches for selected cases. The results for the robot experiments are presented showing quantitative performance figures as well as illustrative on-board sensor snapshots.

5.1 Image Registration Results

Figure 7 presents the quantitative image registration success rates for the three evaluated techniques: SeqSLAM, CNN and SURF. The SeqSLAM algorithm is shown to outperform both the CNN and SURF algorithms in all but one category, matching SURF in one. Of particular relevance to the application domain is the performance of alignment in the multimodal cases (categories 4-11), where the SeqSLAM algorithm is shown to outperform the other methods with 60.0% of alignments considered successful over the 14.2% and 12.9% for the CNN and SURF methods, respectively.

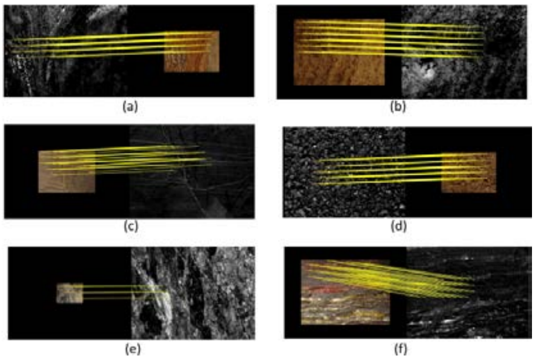

Figure 6: Examples of the feature matches between images made with the SeqSLAM method and used in estimating suitable transforms for image alignment. (a) Case 145, Context 72, Category 11 (b) Case 2149, Context 71, Category 7 (c) Case 3725, Context 96, Category 7 (d) Case 332, Context 78, Category 10 (e) Case 3828, Context 100, Category 6 and (f) Case 4458, Context 108, Category 6.

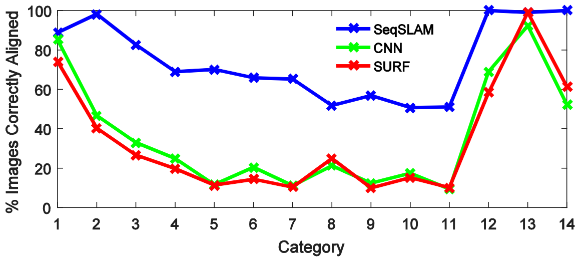

Figure 7: Multimodal rock image dataset registration results. For the three different methods investigated, the percentage of correctly aligned images in each category is shown.

Table 7: Field of View Effects

	SeqSLAM	**Total**
Query image FOV larger	518 (87.2%)	594
Query image FOV smaller	478 (80.5%)	594

With categories 4-7 and 8-11 assessing the same cases but with the images provided in the alternate order, it is shown that the SeqSLAM method is sensitive to which image is used in which stage of the algorithm. Based on the results, the algorithm appears to perform better with the PIXL image with the wider field of view (FOV) of the sample acting as the query image with a success rate of 67.3% as compared to 52.8% with the images swapped. To determine whether this difference in results is due to the mode or the field of view for the query image, an additional breakdown of categories 1 and 3 (the single mode WATSON cases with scale difference) is shown in Table 7. These results suggest that the larger field of view hypothesis is correct.

Figure 6 shows 6 examples of successful image matches made across the different sensing modalities using SeqSLAM, despite large changes in lighting and translational offsets.

While the CNN method is demonstrated to outperform SURF in the multimodal categories, it is expected that given a suitable rock image dataset for training, better performance may be achieved by learning task-specific low-level features. Additionally, use of more recent CNN architectures, use of higher-level features or alternate descriptor methods may improve results. These are left for future work.

5.2 Ideal Lighting Robot Experiment

Through 20 trials, the method was deemed successful 18 times (90% success rate) based on the assessment of a human operator regarding the alignment of the region of interest and the final sensor view. For each trial, Table 8 lists the error in the sensor-surface distance estimate made at the beginning of a trial and the final average pixel distance. The average pixel distance between the region of interest and the sensor's final view is calculated as the average of 4 manually matched points between the images (with matched image width of 960 pixels, and cropped to the minimum height between the two images) and additionally expressed as a percentage of the image width. Success is reported where average pixel distance is less than 10% of the image width. Figure 8 shows an example of the initial stand-off image, manually selected target region of interest and the transition from initial random toolpoint location to the target location.

Table 8: Results of Ideal Lighting Experiments

Trial	Error in Initial Sensor-Surface Distance Estimate (mm)	Success	Average Pixel Distance
1	6.3	Yes	4.7 (0.49%)
2	-15.7	Yes	29.5 (3.08%)
3	6.1	Yes	27.1 (2.82%)
4	11.2	Yes	28.0 (2.92%)
5	-20.1	Yes	11.1 (1.15%)
6	5.1	Yes	7.6 (0.78%)
7	-26.2	Yes	15.6 (1.62%)
8	99.0	Yes	13.3 (1.39%)
9	44.6	Yes	13.1 (1.37%)
10	68.1	Yes	24.4 (2.55%)
11	1.9	Yes	42.5 (4.43%)
12	-66.9	Yes	15.7 (1.64%)
13	-58.5	Yes	33.2 (3.46%)
14	43.1	Yes	45.6 (4.75%)
15	-32.4	Yes	9.0 (0.93%)
16	69.6	**No**	112.7 (11.74%)
17	28.2	**No***	20.6 (2.14%)
18	-79.4	Yes	22.6 (2.35%)
19	36.2	Yes	25.4 (2.65%)
20	-36.6	Yes	35.8 (3.73%)

* Trial 17 failed by moving to an inappropriate initial estimate pose however managed to correctly move to the final position

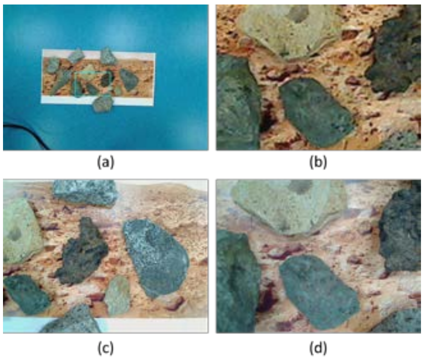

Figure 8: Camera view for (a) initial stand-off image, with (b) manually selected region of interest. (c) View from the random initial toolpoint camera position, and (d) view after visual positioning to target has completed.

5.3 Varied Illumination Robot Experiment

To simulate Mars-like lighting cycles, we used strong artificial lighting to drastically change the lighting conditions during the experiment. The rock samples were rearranged in each trial allowing for the generation of highly textured sample surfaces and subsequently significant shadowing differences and perspective changes between the initial reference image and the sensor's view, as shown in Figure 9.

For this experiment, using SURF image registration for the final fine tuning toolpoint movement failed consistently. Replacing SURF with SeqSLAM enabled the toolpoint to move to the correct final location, despite the large changes in appearance evidenced in Figure 9e. Given the computational cost of the SeqSLAM algorithm (discussed later), PID control was very slow; instead we used an open-loop control mechanism (future work discusses possible optimizations).

Table 9 shows the lighting conditions, the experimental error in the distance estimate, success/failure of the trial and average pixel distance between the two images. Of the 20 trials, the sensor successfully manoeuvred to the region of interest in 13 cases. Figure 10 shows several examples of the operator-selected region of interest and final sensor view; the difference in illumination is quite apparent, significantly altering the appearance of the sample between region selection and the registration stages.

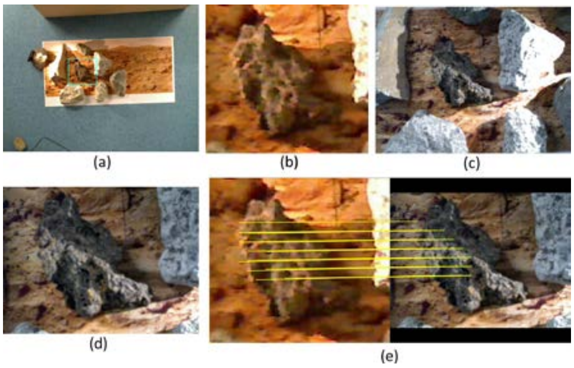

Figure 9: Camera view for (a) initial stand-off image, with (b) manually selected region of interest, with significant lighting changes. (c) View from the random initial toolpoint camera position, and (d) view after visual positioning to target has completed. (e) Feature correspondences found between the stand-off image and the toolpoint camera image at the final location using SeqSLAM.

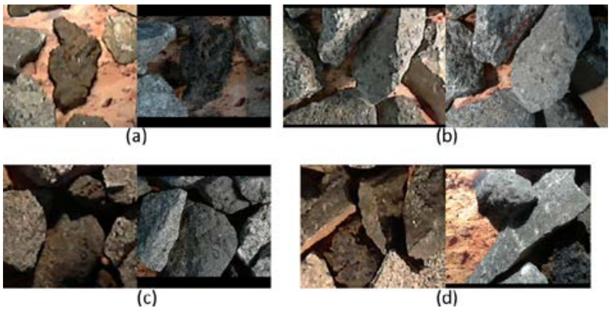

Figure 10: Examples of (left) the selected region of interest and (right) the final sensor view in successful trials 4, 7 and 17 (a-c, respectively) and (d) unsuccessful trial 18. The black bands are present due to the differing aspect ratio of the ROI and the sensors image dimensions.

Table 9: Results of Lighting Difference Experiments

Trial	Error in Initial Sensor-Surface Distance Estimate (mm)	Initial Lighting	Motion Lighting	Success	Average Pixel Distance
1	-61.7	Left	Right	Yes	32.0 (3.3%)
2	2.8	Right	Left	Yes	16.8 (1.8%)
3	26.9	Left	Right	Yes	10.3 (1.1%)
4	75.2	Top	Bottom-Right	Yes	20.0 (2.1%)
5	**113.0**	Left	Top-Right	**No**	Unmeasurable
6	-62.7	Bottom-Left	Right	Yes	22.2 (2.3%)
7	-16.4	Right	Left	Yes	32.7 (3.4%)
8	-107.8	Bottom-Right	Left	Yes	13.6 (1.4%)
9	62.4	Bottom	Top-Right	**No**	Unmeasurable
10	14.9	Top-Right	Bottom-Left	**No**	Unmeasurable
11	48.8	Right	Left	Yes	10.8 (1.1%)
12	-28.4	Top	Bottom-Left	**No**	Unmeasurable
13	**123.0**	Bottom-Left	Right	**No**	Unmeasurable
14	-40.5	Right	Top-Left	Yes	54.1 (5.6%)
15	27.4	Top-Left	Bottom-Left	Yes	14.9 (1.6%)
16	-44.6	Left	Top	Yes	20.3 (2.1%)
17	-181.5	Top	Right	Yes	10.2 (1.1%)
18	**122.3**	Right	Left	**No**	131.6 (13.7%)
19	-178.0	Top-Left	Bottom-Left	Yes	33.9 (3.5%)
20	7.9	Bottom	Top	**No**	515.6 (53.7%)

In failure cases 5, 13 and 18 a large positive error was introduced in the initial estimate of the distance to the surface, resulting in overestimation of sensor-surface distance, overshooting the initial movement towards the region of interest and subsequent image registration attempts failing or returning false-positive registration. Additional factors influencing the failure cases include the significant perspective changes between the region of interest and the sensor's current view in cases where the sample surface is very rough, and the large resolution difference between the region of interest image and that of the sensor image.

5.4 Computational Efficiency

The current algorithms are implemented as unoptimized MATLAB code and run at less than real-time. However, SeqSLAM computation is dominated by the simple patch comparison computation, which is bandwidth rather than computation limited. Consequently, it is possible to calculate reasonably accurate projections of the speed of an optimized implementation by calculating the maximum number of pixel to pixel comparisons required:

$$
\begin{aligned}
B &= s_p \times s_{ROI} \times n_p \times n_s \\
B &= 41^2 \times \frac{300 \times 400}{4} \times 5 \times 150 \\
B &\approx 37.5 \times 10^9
\end{aligned} \tag{5}
$$

where s_p is the patch size in pixels, s_{ROI} is the region of interest image size (estimated as a maximum ¼ of the reference image, often smaller), n_p is the number of patch comparisons performed per image search path and n_s is the number of search paths performed per image.

During the robot experiments, only the initial registration stage would require this number of comparisons. On the assumption that the initial registration and initial positioning estimate are relative accurate, subsequent registration steps may require significantly fewer comparisons based on local region searches, smaller patch size and/or reduced resolution. Based on a similar calculation as above at quarter patch size, local searching in one quarter of image and at one quarter resolution (i.e. roughly the resolution of the original region of interest), only ~0.65×10^9 comparisons are required.

Based on 0.65 billion comparisons per registration, Table 10 provides estimates of the closed loop positioning update rates achievable on various hardware – we are waiting on finalized computational specifications for the sample rover domain but the low-end Snapdragon processor calculations should give an indication of feasibility on embedded hardware.

Table 10: Update Rate Estimates for Optimised Algorithm

Processor	Update Rate (Hz)
Snapdragon 410 (Embedded CPU)	3.7
Intel Core i7-6700K (High-End CPU)	12
Tegra X1 (Mobile GPU)	19
Titan X (High-End GPU)	258

6 Discussion and Future Work

Performance on the large-scale rock dataset was generally high, with the SeqSLAM approach generally outperforming the other feature-based methods. However, several failures during the robot experiments indicate that the current image registration method and/or experimental process requires failure testing stages to avoid positioning the sensor in an incorrect position or colliding with rock outcrops as a result of false-positive image registration. Although a similarity transform was used in this work, there may be situations where affine transformations may improve alignment results and hence improve the accuracy of the visual positioning process.

The initial target application – visual positioning of a robot arm toolpoint on the Mars2020 rover – has enabled us to make the assumption that the relative pose of the two sensors is approximately known, and instead focus on lateral viewpoint change, lighting change, and the challenges of multi-modal image registration. However, this assumption is not necessarily valid in other applications, and hence a logical future step would be to investigate adding scale-invariant feature detection/matching and subsequent image alignment. The results presented here demonstrate that feature correspondences can be found at significantly different scales despite all other challenges – but did not require a search through all possible relative scales.

Although the SeqSLAM method implementation here was far from real-time, the simplicity of its approach should mean that current work developing an optimized version should yield a real-time capable system, even on limited computational hardware; this has already been shown in the navigation domain [Arroyo *et al.*, 2015]. An assessment of the suitability for porting the proposed methods to an embedded system should also be investigated: FPGA's or other hardware capable of parallel computing are potential solutions.

Other alignment techniques such as mutual information should also be investigated, although they will require some adaptation due to their reliance on accurate extrinsic information for calibrated sensors. Sensors may become uncalibrated during Mars operations due to long operation times and extreme environmental conditions. ORB, a combination of oriented FAST and rotated BRIEF features, is another potential solution, with attractive scale-, rotation- and noise-invariant properties [Rublee *et al.*, 2011].

For the CNN method, we are currently performing unsupervised learning of appropriate low-level filters for each of the PIXL and WATSON datasets. It is hypothesised that with filters learned from the datasets, image features of higher significance may be more readily identified. In addition to this, use of increasingly abstract filters and the subsequent deeper level feature maps may increase identification of significant image features. Additional performance gains will probably require techniques such as distance metric learning that learn the specific transformation between sensory modes.

A major practical concern for the Mars mission application is safety. Currently the system does not explicitly address collision avoidance. Potential solutions may include feature point tracking and 3D reconstruction and assessment of the percentage of matched point features used in the estimation of the similarity transform.

In conclusion, the results presented here demonstrate the potential for robust image registration and vision-based positioning across multiple sensing modalities, resolution changes, lighting conditions, significant scale changes (when approximate relative poses are known) and lateral viewpoint changes. We are currently expanding this research towards a range of other application domains, where the key requirement is reliable and accurate positioning of a toolpoint, mobile robot or UAV relative to a surface of interest, whether it be inspection of the metal skin of an aircraft for lightning strike damage or the concrete wall of an industrial plant for cracks.

7 Acknowledgements

We thank Sue Keay for kindly providing the rock samples used in the robot arm experiments. We thank Christian Liebe (Jet Propulsion Laboratory, California Institute of Technology) for assistance in acquiring image data, and the PIXL instrument team for their advice and counsel. Finally, we thank Thom Wynne of the Jet Propulsion Laboratory photography lab for his help in arranging the testbed.

References

[Allwood, A., *et al.*, 2015 of Conference] Allwood, A., *et al.* (2015 of Conference). *Texture-specific elemental analysis of rocks and soils with PIXL: The Planetary Instrument for X-ray Lithochemistry on Mars 2020.* Paper presented at the 2015 IEEE Aerospace Conference. 2015 of Conference

[Arroyo, R., *et al.*, 2015] Arroyo, R., *et al. Towards life-long visual localization using an efficient matching of binary sequences from images.* Paper presented at the 2015 IEEE International Conference on Robotics and Automation (ICRA) 2015.

[Azizian, M., *et al.*, 2014] Azizian, M., *et al.* Visual servoing in medical robotics: a survey. Part I: endoscopic and direct vision imaging – techniques and applications. *The International Journal of Medical Robotics and Computer Assisted Surgery, 10*(3), 263-274. doi: 10.1002/rcs.1531 (2014)

[Bay, H., *et al.*, 2006] Bay, H., *et al. SURF: Speeded Up Robust Features.* Paper presented at the European Conference on Computer Vision 2006.

[Biederman, I., 1988] Biederman, I. Aspects and extension of a theory of human image understanding. *Computational processes in human vision: An interdisciplinary perspective*. (1988)

[Bonin-Font, F., *et al.*, 2008] Bonin-Font, F., *et al.* Visual navigation for mobile robots: A survey. *Journal of Intelligent & Robotic Systems, 53*(3), 263-296. doi: DOI 10.1007/s10846-008-9235-4 (2008)

[Corke, P., 2011] Corke, P. (2011). *Robotics, Vision and Control: Fundamental Algorithms in MATLAB* (Vol. 73): Springer.

[Corke, P., *et al.*, 2013] Corke, P., *et al. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation.* Paper presented at the Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on 2013.

[Corke, P. I., 1996] Corke, P. I. (1996). *Visual Control of Robots: high-performance visual servoing*: Research Studies Press Baldock.

[Johns, E. & Yang, G. Z., 2013] Johns, E. & Yang, G. Z. *Feature Co-occurrence Maps: Appearance-based Localisation Throughout the Day.* Paper presented at the International Conference on Robotics and Automation, Karlsruhe, Germany 2013.

[Kanji, T., 2015] Kanji, T. *Cross-season place recognition using nbnn scene descriptor.* Paper presented at the Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on 2015.

[Kermorgant, O. & Chaumette, F., 2011] Kermorgant, O. & Chaumette, F. *Multi-sensor data fusion in sensor-based control: application to multi-camera visual servoing.* Paper presented at the Robotics and Automation (ICRA), 2011 IEEE International Conference on 2011.

[Kosmopoulos, D., 2011] Kosmopoulos, D. Robust Jacobian matrix estimation for image-based visual servoing. *Robotics and Computer-Integrated Manufacturing, 27*(1), 82-87. (2011)

[Kragic, D. & Christensen, H. I., 2002] Kragic, D. & Christensen, H. I. Survey on visual servoing for manipulation. *Computational Vision and Active Perception Laboratory, Fiskartorpsv, 15*. (2002)

[Krizhevsky, A., *et al.*, 2012] Krizhevsky, A., *et al. Imagenet classification with deep convolutional neural networks.* Paper presented at the Advances in Neural Information Processing Systems 2012.

[Leeper, A., *et al.*, 2014] Leeper, A., *et al. Using near-field stereo vision for robotic grasping in cluttered environments.* Paper presented at the Experimental Robotics 2014.

[Lowry, S., *et al.*, 2014] Lowry, S., *et al. Transforming Morning to Afternoon using Linear Regression Techniques [under review].* Paper presented at the IEEE International Conference on Robotics and Automation, Hong Kong, China 2014.

[Maddern, W., *et al.*, 2014] Maddern, W., *et al. Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles.* Paper presented at the Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China 2014.

[Malis, E. & Rives, P., 2003] Malis, E. & Rives, P. Robustness of image-based visual servoing with respect to depth distribution errors. … . *ICRA'03. IEEE International Conference on*. (2003)

[Mehta, S. S. & Burks, T. F., 2014] Mehta, S. S. & Burks, T. F. Vision-based control of robotic manipulator for citrus harvesting. *Computers and Electronics in Agriculture, 102,* 146-158. doi: 10.1016/j.compag.2014.01.003 (2014)

[Michaels, A., *et al.*, 2015] Michaels, A., *et al. Vision-based high-speed manipulation for robotic ultra-precise weed control*. Paper presented at the Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on 2015.

[Milford, M., 2013] Milford, M. Vision-based place recognition: how low can you go? *International Journal of Robotics Research, 32*(7), 766-789. (2013)

[Milford, M., *et al.*, 2014a] Milford, M., *et al. Automated sensory data alignment for environmental and epidermal change monitoring.* Paper presented at the Australasian Conference on Robotics and Automation 2014 2014a.

[Milford, M., *et al.*, 2014b] Milford, M., *et al. Condition-Invariant, Top-Down Visual Place Recognition.* Paper presented at the IEEE International Conference in Robotics and Automation (ICRA) 2014b.

[Milford, M. & Wyeth, G., 2012] Milford, M. & Wyeth, G. *SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights.* Paper presented at the IEEE International Conference on Robotics and Automation, St Paul, United States 2012.

[Naseer, T., *et al.*, 2014] Naseer, T., *et al. Robust visual robot localization across seasons using network flows.* Paper presented at the Conference on the Association for the Advancement of Artificial Intelligence 2014.

[Neubert, P., *et al.*, 2013 of Conference] Neubert, P., *et al.* (2013 of Conference). *Appearance change prediction for long-term navigation across seasons*. Paper presented at the European Conference on Mobile Robots, Barcelona, Spain. 2013 of Conference

[Novak, K. S., *et al.*, 2015] Novak, K. S., *et al. Preliminary Surface Thermal Design of the Mars 2020 Rover* 2015.

[Pepperell, E., *et al.*, 2014] Pepperell, E., *et al. All-environment visual place recognition with SMART.* Paper presented at the Proceedings of the International Conference on Robotics and Automation 2014.

[Ranganathan, A., *et al.*, 2013] Ranganathan, A., *et al. Towards illumination invariance for visual localization.* Paper presented at the Robotics and Automation (ICRA), 2013 IEEE International Conference on 2013.

[Rublee, E., *et al.*, 2011] Rublee, E., *et al. ORB: An efficient alternative to SIFT or SURF.* Paper presented at the 2011 International conference on computer vision 2011.

[Sa, I., *et al.*, 2015] Sa, I., *et al.* Inspection of pole-like structures using a visual-inertial aided VTOL platform with shared autonomy. *Sensors (Switzerland), 15,* 22003-22048. doi: 10.3390/s150922003 (2015)