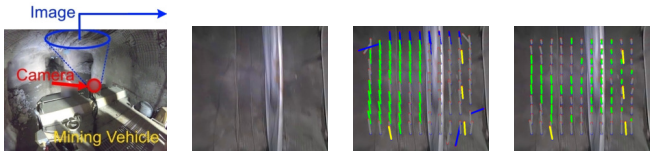
I2-S2: Intra-Image-SeqSLAM for More Accurate Vision-based Localisation in Underground Mines

Fan Zeng[1], **Adam Jacobson**[1], **David Smith**[2], **Nigel Boswell**[2], **Thierry Peynot**[1], **Michael Milford**[1]
[1]Queensland University of Technology, Australia
[2]Caterpillar, Inc.
[1]fan.zeng@qut.edu.au *

Abstract

Many real-world robotic and autonomous vehicle applications, such as autonomous mining vehicles, require robust localisation under challenging environmental conditions. Laser range sensors have been used traditionally, but often get lost in long tunnels that are the major components of underground mines. Recent research and applied systems have been increasingly using cameras, bringing in new challenges with regards to robustness against appearance and viewpoint changes. In this paper we develop a novel visual place recognition algorithm for autonomous underground mining vehicles that can be used to provide sufficiently accurate (sub-metre) metric pose estimation while also having the appearance-invariant and computationally lightweight characteristics of topological appearance-based methods. The challenge of large viewing angle variations typical in confined tunnels is addressed by incorporating multiple reference image candidates. The framework is evaluated with real-world multi-traverse datasets featuring different environments including underground mining tunnels and office building environments. The reprojection error of image registration is ∼50% lower than a state-of-the-art deep-learning based method (MR-FLOW) using manually-labelled ground truth on a set of images representing typical scenarios during the underground mining process.

*This research was supported by an Advance Queensland Innovation Partnerships grant from the Queensland Government, Mining3, Caterpillar and the Queensland University of Technology. MM also received support from an ARC Future Fellowship FT140101229.



(a) (b) (c) (d)

Figure 1: (a) Picture of a mining vehicle in an underground tunnel, with vision sensors mounted to capture images of the ceiling. (b) Example of a captured ceiling image. (c) The image in (b) overlapped with sparse optical flow w.r.t a reference image (not shown) from the proposed I2-S2 algorithm. (d) Corresponding optical flow field extracted from the dense result calculated by a state-of-the-art deep-learning algorithm (MR-FLOW), underestimating the velocity of the vehicle. Optical flow vectors are colour coded as follows: Yellow - manually labelled ground truth; green - accepted optical flow vectors; gray and blue - discarded optical flow vectors.

1 INTRODUCTION

Mining vehicles have been improving production efficiencies in modern underground mines. Though some of them are autonomous to some extent, there is significant industry pressure to further improve the level of autonomy by enhancing their localisation capability. Existing techniques, such as the Wireless Sensor Networks (WSN) [Moridi *et al.*, 2015] and Radio-Frequency IDentification (RFID)-based [Lavigne and Marshall, 2012] underground localisation, inevitably require extensive surveying and investment on the infrastructure. In contrast, the cost-effective vision-based localisation is becoming more attractive with the development of computer vision technologies.

Current approaches for this purpose approximately fall into three broad categories: i) Very accurate, full 3D / 6DOF (Degrees of Freedom) visual SLAM or positioning, ii) topological appearance-based methods and iii) LiDAR-based underground vehicle localisation methods that are neither full 3D SLAM nor topological, e.g. [Daoust *et al.*, 2016]. The full 3D solutions typically

require significant computation and are sensitive to extreme appearance changes, while the topological techniques do not provide any or sufficient metric positioning information. A light-weight camera-only coarse localisation system for underground environment [Jacobson *et al.*, 2018] [Zeng *et al.*, 2017] has been developed by the authors up to a localisation accuracy of ∼9 metres. To ultimately achieve fully-automated underground mining process, it is necessary to further pin-point the location of the vehicle within sub-metre accuracy, especially w.r.t digging / tipping points, so as to better assist obstacle and pedestrian avoidance by knowing with high degree of certainty the accurate vehicle locations relative to obstacles and pedestrians captured by other sensors on the vehicle.

The underground mining context offers unique opportunities and challenges. On one hand, the constituent objects and patterns in images from a mine tunnel are relatively predictable, allowing an application-specific design methodology. On the other hand, repetitive patterns, and the presence of dust and water increase the difficulty in precise alignment, in addition to large relative displacements when the vehicle is moving. Variable lighting conditions, especially the lack of ambient light is another major challenge for underground localisation. Most real operating mines have no lights, except perhaps in some limited areas. Some mines are lit in places, but even then the amount of lighting is significantly less than on surface. Even with active lighting added, fast moving vehicles in comparatively dark conditions results in a lot of motion blur, which is a huge issue for cameras that move underground. Last but not least, the field of view (FOV) is often limited due to the lower height of mine tunnel ceilings, evident even with fish-eye cameras: The same scene can appear dramatically different when viewed from locations found to be close by coarse localisation, to the extent that it becomes impossible to directly solve the sub-metre localisation problem with a single camera-based reference map. To resolve this issue, a scheme with multiple reference images covering a larger area that is equivalent to the expansion of FOV is introduced in this paper.

An algorithm based on intra-image patch-sequence matching that is robust to appearance change is presented in this paper. The relative displacement between the query and a set of reference locations / poses is obtained by matching selected point pairs in images (Fig. 1c). The query and reference images are taken by a fish-eye camera mounted on the vehicle pointing towards the ceiling (Fig. 1a). Compared with other possible locations to set-up a camera, such as pointing to the front of the vehicle or to the walls, mine tunnel ceilings most consistently offer a quasi-planar surface for homography extraction. Moreover, moving foreground objects, for which modern algorithms tend to devote time to identify, are rare on mine tunnel ceilings, allowing a sparse instead of dense optical flow estimation to suffice.

Main contribution of the paper: 1. Given a pair of query and reference images, a robust algorithm that generates a sparse optical flow field between them for the subsequent homography estimation step to refine the coarse localisation result, is described. 2. A reference select module that pairs up the query image with multiple reference images in order to resolve the limited FOV problem is proposed and verified on datasets collected in different environments (office and mine tunnel).

The paper proceeds as follows. Section 2 reviews previous works on confined space localisation and relevant literature on optical flow based image registration algorithms. Section 3 elaborates the detailed implementations of the proposed approach. Section 4 gives the configurations of experimental set-up to verify the effectiveness of our approach on datasets collected from underground mine tunnels and university lab office environment, with the results of experiments presented in Section 5. A brief discussion and conclusions can be found in Section 6 and 7.

2 Literature Review

2.1 General SLAM and Localisation

3D point clouds from LiDARs have been used in SLAM algorithms [Zhang and Singh, 2015]; laser-based localisation in a mine [Zlot and Bosse, 2014] has also been studied. However, lasers tend to get lost in long uniform tunnels. On the other hand, cameras collecting visual information can work regardless of whether the vehicle is in a long tunnel, but loop closure uncertainty makes monocular SLAM problem difficult. Thanks to the fact that once a high-quality reference database is built for a mine tunnel, one can benefit from the map by traversing the tunnel thousands of times, pre-mapping the mine tunnels before the localisation run is economical. This helps to remove the ambiguities of unfamiliar scenes in regular SLAM problems.

2.2 Vision-based Localisation Techniques

Full 3D / 6DOF (Degrees of Freedom) visual SLAM systems such as ORBSLAM [Mur-Artal Raúl and Tardós, 2015] and LSD-SLAM [Caruso *et al.*, 2015] do not work well in environments with sudden big changes in appearance. The classical probabilistic SLAM system FAB-MAP [Cummins and Newman, 2008] achieves some degree of viewpoint-invariance by associating places with sets of features. Since geometrical information is not retained in its bag-of-words representation, it does not work well under severe visual-aliasing. SeqSLAM [Milford and Wyeth, 2012] has been shown to be more successful by matching sequences of images based on Sum

of Absolute Difference (SAD) scores and using patch-normalisation to minimise the effect of appearance-change. However, as an approach that performs place recognition, it does not provide the metric information required for sub-metre localisation. Our previous work [Zeng *et al.*, 2017] demonstrated vision-based coarse localisation of vehicles inside mine tunnels, but the best match location of whole-image matching does not provide metric indication of pose change. There is a gap between the full-metric and pure-topological approaches, which sub-metre underground localisation calls for.

2.3 Image Registration Techniques

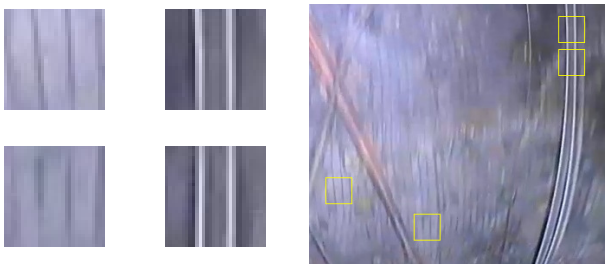
Classical image registration algorithms such as Lucas-Kanade [Lucas *et al.*, 1981] have been widely used to accurately estimate the sub-pixel flow vectors. When the aim comes to getting the relative pose, the flow vector could be either sub-pixel or hundreds of pixels, making the problem difficult. Methods that derive from Lucas-Kanade [Birchfield, 2007] for large displacements are susceptible to visual aliasing. Methods that match interest points (selected by e.g. SURF [Bay *et al.*, 2006] feature detectors) with nearest point matchers such as FLANN [Muja and Lowe, 2014] exploit the advantages of scale and rotation invariance, but again they fall short in visually aliased environments. Recent research have focused on utilising Convolutional Neural Network (CNN) for object recognition [Dosovitskiy *et al.*, 2015] and separating fast-moving objects from the rigid background, among which MR-FLOW [Wulff *et al.*, 2017] has shown robustness with appearance change and motion blur. The results obtained from MR-FLOW without the computation time constraint will be used to benchmark our sub-metre localisation approach under various adverse underground conditions. Nevertheless, it provides more information than needed for our localisation problem because obtaining the dense optical flow for every pixel is unnecessary for homography estimation in environments like a mine tunnel.

3 Approach

Our vision-based sub-metre localisation approach is an extension of the method in [Sergeant *et al.*, 2016] that relies solely upon the matching of RGB camera images. A query image I_{query} of size $Row(I_{query}) \times Col(I_{query})$ of mine tunnel ceiling and a reference image I_{ref} taken from a nearby perspective (∼3-metre neighbourhood) is given as input for relative pose extraction.

3.1 Simple Patch Matching

The base method underlying our novel contribution is introduced first in this subsection. Since moving objects in the foreground are not common in mine tunnel ceiling footages, the visual displacement can be estimated using sparse sample points instead of registering every pixel. This can be confirmed by the resultant near-uniform optical flow field around each sampled point if dense registration is indeed performed. If I_{query} and I_{ref} are known to have some overlap, such that the distance in pixel space between the same feature in the two images is less than a known upper-bound L_{SR}, the best match for a pixel in I_{query} can potentially be found by scanning every candidate pixel in a bounding box of $B_{SR} = [2L_{SR} \times 2L_{SR}]$ in I_{ref}. Though matches can be found in ideal conditions, there are several failure mechanisms. First, repetitive patterns and monotonic feature-sets in a mine tunnel environment can create severe visual aliasing within the image, as shown in Fig. 2. The two image patches in Fig. 2a as well as Fig. 2b are cropped from different locations of the same image (Fig. 2c, yellow boxes), despite the nearly identical looks. Confusing them when picking the best-match image patch will lead to errors in sub-metre localisation. Increasing patch size normally would partially resolve the visual aliasing by resorting to finer details, at the cost of computation, as well as requiring more expensive cameras to provide higher resolution images. However, it does not work in a mining context because images are often dusty and blurry, and there are not much finer details to resort to.



(a) (b) (c)

Figure 2: Visual aliasing in mine tunnel ceiling images. Visually similar patches in (a) and (b) actually come from different locations in (c), as indicated by the yellow boxes.

3.2 Intra-Image-SeqSLAM (I2-S2)

An effective method in addressing visual aliasing for place recognition is SeqSLAM [Milford and Wyeth, 2012]. The same technique of matching a sequence of images can be applied within an image to recognise different intra-image "places", alleviating visual aliasing. The Intra-Image-SeqSLAM (I2-S2) approach generalises SeqSLAM in pixel space within images. In I2-S2, a query "place" $p_{query} = (x_{query}, y_{query})$ is a pixel located within the query image I_{query}, and a reference "place" $p_{ref} = (x_{ref}, y_{ref})$ is a pixel within the reference image

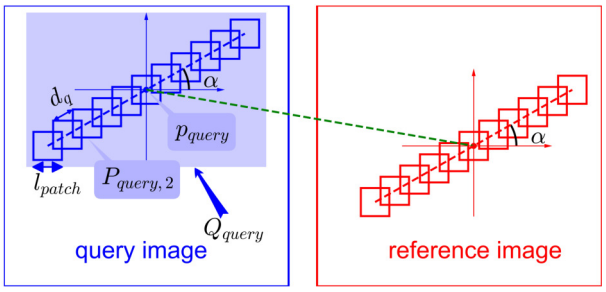


Figure 3: Schematic Diagram of I2S2. The various components are labelled for the query image, the corresponding labels in the reference image are omitted.

I_{ref}. Unlike the vanilla approach that makes matching decisions solely based on the SAD score between the query patch $P_{query}(p_{query})$ centred at p_{query} and every candidate reference patch $P_{ref}(p_{ref})$ centred at p_{ref}, a sequence Q_{query} of patches $P_{query,i}$ centred at a set of points $p_{query,i} \in S(p_{query})$ sampled around the neighbourhood $\delta(p_{query})$ of p_{query} are compared with their counter parts in Q_{ref}, i.e.,

$$Q_X = \{P_{X,i}(p_{X,i}) | p_{X,i} \in S(p_X)\}, X \in (query, ref). \quad (1)$$

By matching a sequence of patches around the pixel neighbourhood, more contextual information around that pixel is integrated, reducing the visual aliasing effect.

3.3 Sequence Generation Module

A sequence generation module is constructed to handle the many ways to produce a sequence of patches $P(p)$ around a pixel p. It takes in an image and a point of interest p, and generates a sequence Q with length l_q of patch-normalised patches sized $l_{patch} \times l_{patch}$ in the neighbourhood of p. Additional parameters are used to define the specific manner in which the sequence of patches are sampled: The distance (step length) d_q between the adjacent sampled patches, and the angle α of the trajectory line the sequence is sampled along. An example of sequences Q_{query} and Q_{ref} is shown in the schematic of I2-S2 in Fig. 3. Note that although the patch at the centre of the sequences is centred at p_{query} and p_{ref} in the schematic, this is not required since I2-S2 only cares about the neighbourhood as a whole. In fact, p_{query} and p_{ref} have no more weight assigned to them than other pixels in the sequence. For those p_{query} and p_{ref} near the image edge, d_q is reduced to "squeeze-in" the part of generated sequences that goes beyond the image boundary in our current implementation. This flexibility in sampling the sequence may distort the relative correspondence between patches in Q_{query} and those in Q_{ref}, therefore a novel way to calculate the sequence score is introduced.

3.4 Sequence Score

After sequence Q_{query} is generated around p_{query}, along with one $Q_{ref}(x, y)$ for each reference candidate $p_{ref}(x, y)$, where $(x, y) \in B_{SR}$, a confusion matrix $M_{conf,(x,y)} \in \mathbb{R}^{l_q \times l_q}$ can be obtained for each $Q_{ref}(x, y)$, after each pair $(P_{query} \in Q_{query}, P_{ref} \in Q_{ref})$ of patches is compared, as in original SeqSLAM. A trajectory $T = \{(i \mapsto j = T(i)) | i, j \in \mathbb{N}, 0 \leq i, j \leq l_q\}$ is a mapping between P_{query} in Q_{query} and P_{ref} in $Q_{ref}(x, y)$. A coherent trajectory $\tilde{T}$ is one that satisfies $\tilde{T}(i_1) \leq \tilde{T}(i_2)$, if $i_1 < i_2$. The sequence score $E(p_{ref}(x, y))$ for each reference candidate $p_{ref}(x, y)$ is the minimum trajectory score for all coherent trajectories on the confusion matrix $M_{conf,(x,y)}$, i.e.

$$E(p_{ref}(x, y)) = \min_{\tilde{T}} \sum_{(i \mapsto j) \in \tilde{T}} M_{conf,(x,y)}(i, j) \quad (2)$$

which can be found in $\Theta(l_q^2)$ with dynamic programming. Effectively, this downweighs the match between the patch centred exactly at p_{query} and p_{ref}, and blend in the best possible contribution from their neighbourhood. Finally, the matching reference point $p_{ref}(x^*, y^*)$ is the one with the minimum sequence score:

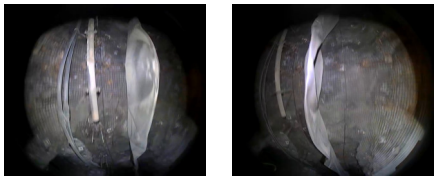
$$(x^*, y^*) = \arg\min_{(x,y)} (E(p_{ref}(x, y)). \quad (3)$$

3.5 Homography

For each pair of query and reference images, a Manhattan grid G of size $h_G \times w_G$ of sampling points l_G pixels apart is used to extract a series $V(p_{query})$ of query points. For each $p_{query} \in V(p_{query})$, a range B_{SR} centred at p_{query} to search for matching reference point is defined. A series $V(p_{ref})$ of points from the reference image that best matches $V(p_{query})$ is obtained using I2-S2. After a preliminary filtering of the "out-of-range" matches (those with matching p_{ref} on the boundary on B_{SR}), the mapping $MAP(V(p_{query}), V(p_{ref}))$ is fed into a RANSAC filter before a homography $H(\hat{V}(p_{ref}), \hat{V}(p_{query}))$ is calculated based on the filtered $MAP(\hat{V}(p_{ref}), \hat{V}(p_{query}))$. The homography H can subsequently be used to estimate vehicle's relative pose transform between the query and reference images.

3.6 Algorithmic Complexity

One can derive from the descriptions above the asymptotic complexity for generating $MAP(V(p_{ref}), V(p_{ref}))$ is $\Theta(h_G w_G L_{SR}^2 l_q^2 l_{patch}^2)$. The need to find a match for every pixel in the query images is relaxed to a sparse grid G whose density can be adjusted according to the density of non-planar objects on the ceiling. Furthermore, the size and sparsity of the search box B_{SR} can also be tuned according to maximum possible vehicle velocity for further reduction in computation time. Though



(a) (b)

Figure 4: Images of the same place taken from two perspectives. The same ceiling pipelines in (a) also appear in (b), shifted towards the left. Dramatic visual changes significantly degrades the performance of general visual matching methods. Note the images are already taken using a fish eye camera.

sequence length l_q and patch size l_{patch} both have super-linear influence in terms of complexity, in practice they seldom need to be set too large, as long as the sequence of patches generated by l_q, l_{patch} as well as step length d_q spans a neighbourhood distinguished enough to identify the pixel "place" with confidence.

3.7 Reference Select Module

In practice, even if coarse place recognition correctly provides an I_{ref} taken from a nearby location of I_{query}, the relative pose change could be so large that there is little overlap between the images. Even when there is overlap, viewing the same scene from different angles can bring about dramatic visual changes such as those shown in Figure 4, which challenges most vision-based methods. Moreover, even if matching features do coexist in images I_{query} and I_{ref}, there could be large separation between them that requires a large L_{SR} of the search box B_{SR} to encompass this maximum distance, which quadratically slows down computation.

An effective augmentation to make our method robust to such large pose change is recording a reference database from various viewing angles and perform I2-S2 w.r.t. multiple nearby reference images from those perspectives. After multiple images of different perspectives for a same location are recorded, the query images are then compared with these references until a clear match is found. A reference select module is integrated into the system to select the best reference image. In this way, the size L_{SR} of the search box B_{SR}, which has quadratic influence on complexity, can be constrained, while the cost for the additional reference matching is linear, and can be stopped once a confident match is found, the computation time is reduced. Better accuracy can also be obtained by combining the results from various reference images.

In the next section, the datasets and experimental setup for evaluating the image registration performance of I2-S2 benchmarked by a state-of-the-art deep-learning based approach is described, along with settings for verifying the functionalities of the reference select module.

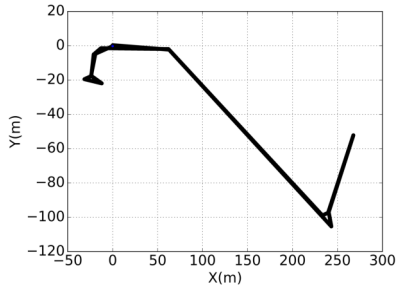


Figure 5: Floor plan of the mine tunnel traversed.

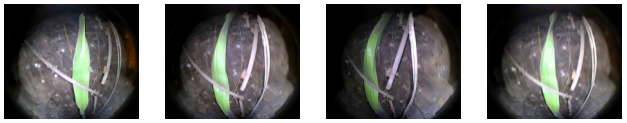
4 Experimental Setup

4.1 Datasets

Three datasets were used to evaluate the I2-S2 performance in this work. The source images of these datasets were collected from an underground mine and inside an office building. The details of these datasets follows:

Mine tunnel multi-traverse dataset
The images were taken using a rugged monocular camera with a 180 degree FOV and image resolution of 944 × 800, mounted on the roof of a light vehicle, pointing towards the ceiling of a mine tunnel. The dataset was collected when the light vehicle travelled along a mine tunnel, the schematic floor plan of which is shown in Fig. 5. The same mine tunnel was traversed four times. During first three traverses, the reference database, which contains three sets of images ("left" , "middle" and "right"), was collected. We drove as much to the left, middle, and right of the road as possible when collecting the "left", "middle" and "right" sets of reference image, respectively. During the fourth traverse, the query images were collected. We drove deliberately in a "zigzag" pattern to increase the difficulty of localisation. Each of the four traverses contains ∼5000 frames as the vehicle goes from the start point through the tunnel, turns around and drives back. A sample frame consisting four images taken at nearby locations during four traverses is shown in Fig. 6. The relative road positioning of the vehicle is evident in the figure from each traverse.

Mine tunnel representative dataset
This dataset consists of 11 triplets of images selected from the "middle" traverse of the "mine tunnel multi-traverse dataset" to be analysed in details. The images were selected to reflect typical conditions in the mining context: Ventilation pipelines, meshes, rock bolts, compromised with dust and motion blur. Each triplet of images consists of 3 consecutive frames, the time interval between adjacent frames is ∼ 70 milliseconds, corresponding to a frame rate of ∼ 14 fps (frames per second). The triplets are independent of each other and are ordered simply based on the time stamp.



(a) (b) (c) (d)

Figure 6: A sample frame from the "mine tunnel multi-traverse dataset", which consists of images from traverses (a) "left", (b) "middle", (c) "right", and (d) "zigzag".

Office ceiling dataset

This dataset features synchronised lab ceiling image sequences captured by four cameras with $\sim 100°$ FOV and a resolution of 640 × 480. A square path in lab area with ceiling ∼1.5 metre above the cameras were travelled twice. Three of the cameras (Cam_L, Cam_M, Cam_R) were fixed to the left, middle and right of a metal bar, with ∼44 cm separation to collect reference images. The fourth camera (Cam_Q) was used to collect query images, it was fixed first between Cam_L and Cam_M for the first loop of travel (frames 0 to 799), then between Cam_M and Cam_R for the second loop (frames 800 to 1599). The FOV of all cameras were cropped to 200 × 200 pixels, and an additional back-and-forth motion (switching direction every 130 frames) was applied to the 200 × 200 crop window of Cam_Q to simulate zigzag motion.

4.2 Preprocessing

The barrel distortions in mine tunnel ceiling images taken by the fish-eye camera were compensated and cropped to 544 × 500 pixels, converted into 8-bit gray-scale, down-sampled 272 × 250 pixels for use in I2-S2.

4.3 Ground truth

Mine tunnel multi-traverse dataset

To build ground truth for evaluation of the reference select module, images from the four traverses were synchronised into frames that contain one image from each traverse ("left", "middle", "right", and "zigzag"). Each frame of four synchronised images were manually examined to see if one of the three reference traverses exhibits an overwhelmingly better viewing-angle resemblance to the query image from the "zigzag" traverse. If yes, the most similar traverse was marked as the ground truth; if no - most likely because the vehicle was indeed at approximately the same location when at least two of the reference images were taken - no ground truth was marked for that particular frame.

Mine tunnel representative dataset

For each pair of images, four patches (P_{query}) that contain distinctive features suitable for manual alignment were selected and $MAP(V(p_{query,gt}), V(p_{ref,gt}))$ was manually aligned. The homography corresponding to this MAP was used as the ground truth.

Office ceiling dataset

The ground truth for evaluation of reference selection module was manually labelled in a similar manner as that of "mine tunnel multi-traverse dataset".

4.4 Benchmark against State of the Art

The open-source Python implementation of MR-FLOW [Wulff *et al.*, 2017] was used to benchmark I2-S2's image registration performance on the "mine tunnel representative dataset". Each triplet of 544 × 500 colour images were fed into MR-FLOW, which output the optical flow between the 2nd and 3rd images. In order to give the algorithm every reasonable advantage possible, MR-FLOW was supplied with full initialisation (including four initial flow fields) except for the "pre-computed rigidity map", which is not available. The four initial flow fields were obtained by running MR-FLOW on corresponding triplets of images without initialisation. Note that the full-fledged initialisation essentially granted MR-FLOW five image frames around the queried frame, thus no recursive effort to improve those initial fields was made. In contrast, the only images provided to I2-S2 were I_{query} (2nd image) and I_{ref} (3rd image), with no initialisation stage. The built-in optimisation was turned on for all MR-FLOW runs.

4.5 Parameters

The parameters used for I2-S2 on the "mine tunnel representative dataset" was summarised in Table 1. A slightly different set was used on "office ceiling dataset" due to the difference in image and general feature sizes.

Table 1: PARAMETER LIST

Parameter	Value	Description
$Row(I)$	250 pixels	downsampled row number
$Col(I)$	272 pixels	downsampled column number
L_{SR}	70 pixels	half size of search box
l_q	15	sequence length
d_q	5 pixels	sample sequence step length
l_{patch}	20 pixels	patch size
l_G	20 pixels	sample grid spacing
$\Delta\alpha$	30 degrees	incremental sequence angle

5 Results

In this section, image registration performance based on sparse optical-flow-field from simple patch matching (subsection 3.1), I2-S2 and the benchmark MR-FLOW is presented and compared. Simple patch matching, henceforth the "vanilla" approach, is simply I2-S2 with sequence length l_q set to 1. The sub-metre accuracy of I2-S2 is first shown. Next, the reference select module is demonstrated to be effective in addressing the problem

of limited FOV. Finally, I2-S2 is shown to be compatible with real-time applications.

5.1 Image Registration Performance

The image registration performance on the "mine tunnel representative dataset" is shown in Fig. 7. The optical flow vectors sampled at an 11×11 grid are plotted for the three methods. The optical flow vectors in blue were considered "out-of-range" by the vanilla method and I2-S2 and discarded before the remaining vectors entered the same OpenCV based RANSAC post-processor for out-lier filtering. Green and grey vectors correspond to the in-liers and out-liers labelled by the RANSAC algorithm, respectively. The homography is calculated using the in-liers only. Manually labelled ground truth vectors are coloured yellow.

Images in the dataset have been selected to represent typical mine tunnel ceiling appearance. Specifically, the 11 triplets cover: Forward motion - Fig. 7(b-d, h-l); backward motion - Fig. 7(g); rotation - Fig. 7(e); idle with flying dust - Fig. 7(f); slow (Fig. 7(g, h)) and fast (Fig. 7(b-d, i-l)) movement; with (Fig. 7(b-j, l)) and without (Fig. 7(k)) various combinations of large objects on the ceiling.

From the optical flow fields, all three methods performed well for the "rotation", "idle with flying dust" and "slow movement" cases. It is worth noting that despite the fact that only coherent correspondence between patches in Q_{query} and Q_{ref} was enforced, and comparing sequences of patches in I2-S2 treated the query pixel p_{query} and reference candidate pixel $p_{ref}(x, y)$ no differently from other pixels, those exact pixels could still be precisely matched, as shown by the idle case. MR-FLOW underestimated the optical flow for the fast movement cases (Fig. 7(d, j, l)). Prominent objects on the ceiling affected the local optical flow for all three methods, with the vanilla method most susceptible. The optical flow field from I2-S2 in the neighbourhood of large ceiling objects were significantly better than the vanilla method, resulting in more in-liers after subsequent filtering, thanks to the effect of sampling a sequence of patches around the query pixel to blend in a larger context surrounding the ceiling object. The blue optical flow vectors near the top of the images in the forward moving cases were labelled "out-of-range" and discarded because the best-match is on the boundary of search range B_{SR}. These were correct decisions since the corresponding pixels those query pixels match to indeed went out of scope in the next frame as a result of the vehicle movement. The in-liers determined by the RANSAC algorithm were used to calculate the homography, the accuracy of which were evaluated in the following way: A set of four $p_{query,test}$ are obtained by projecting the four $p_{ref,gt}$ using the homography, and the reprojection

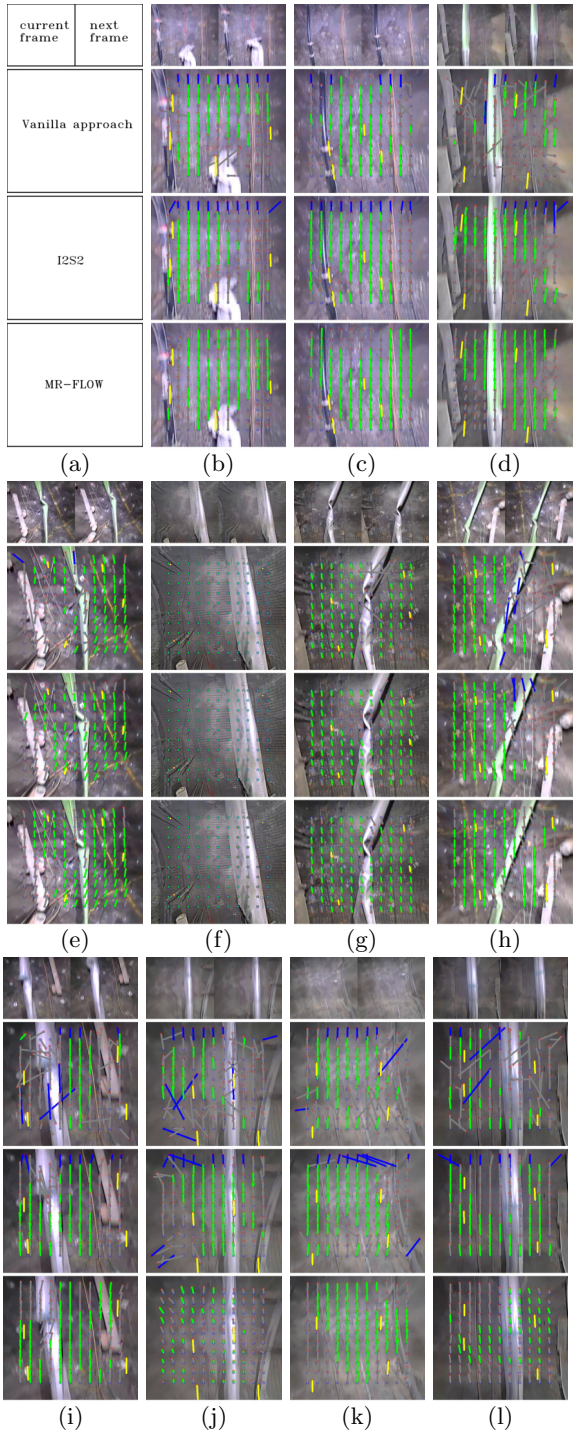


Figure 7: (a) Annotation for subplots in the rest of this figure. First row left: 2nd frame in triplet, right: 3rd frame in triplet. Row 2 to 4: Results from the vanilla approach (subsection 3.1), I2-S2, and MR-FLOW. (b-h) Estimated optical flow field on the "mine tunnel representative dataset". Optical flow vectors are colour coded as follows: Yellow - manually labelled ground truth; green - in-liers after RANSAC; gray - out-liers after RANSAC; blue - considered out of range by the vanilla method and I2-S2.

error d_{reproj} defined below is used as the metric for comparison:

$$d_{reproj} = \left(\frac{1}{|V|} \sum_{p_{query,gt} \in V} (p_{query,test} - p_{query,gt})^2 \right)^{\frac{1}{2}}, \tag{4}$$

in which $V = V(p_{query,gt})$. The reprojection error for the 11 image triplets shown in Fig. 7 is plotted in Fig. 8.

On average, I2-S2 has ∼50% lower reprojection error than MR-FLOW, and ∼ 21% lower than the vanilla approach. Differences in reprojection error that are larger than the estimated error (∼ 2 pixels) in the manually aligned ground truth are significant. For cases (c)(e)(f)(g) the performance for these three methods is similar because the difference is below 2 pixels. For the remaining cases, I2-S2 performs better than the vanilla approach. I2-S2 also shows better performance than MR-FLOW except for case(i), where the results are comparable, and case (h), in which case I2-S2 correctly estimated the optical flow on the right to be larger in magnitude than those on the left, leading RANSAC filter to consider them as out-liers, whereas the more uniform (but less accurate) flow field from MR-FLOW did not have this effect. This directly results from their corresponding accuracy of sparse optical flow field in Fig. 7.

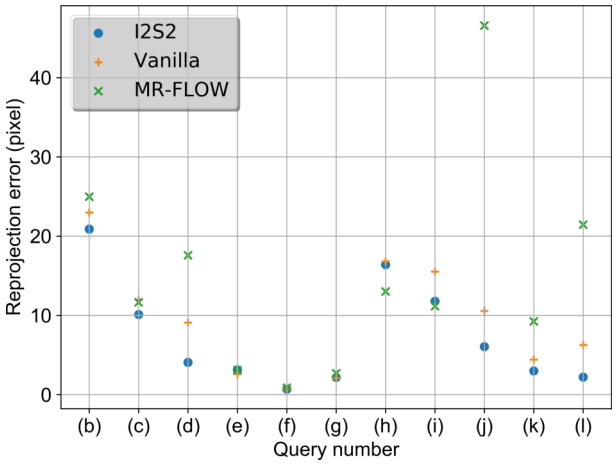
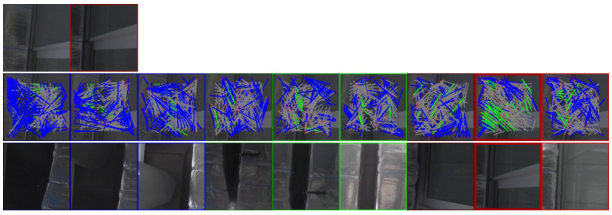
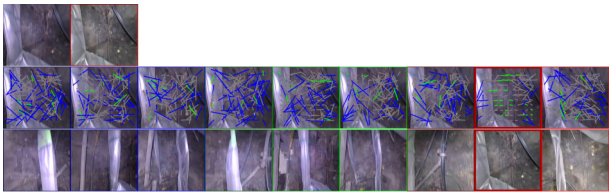


Figure 8: Reprojection error on the manually labelled "mine tunnel representative dataset" by the vanilla approach, I2-S2 and the benchmark algorithm MR-FLOW. The query number corresponds to those in Fig. 7.

The average reprojection error of I2-S2 on the "mine tunnel representative dataset" is ∼7 pixels, which corresponds to approximately 1.4% of the image size (544 × 500). Since the FOV is estimated to be ∼3×3 metres, 1.4% error corresponds to an accuracy of ∼4 centimetres. Even if we use the maximum error of 21 pixels (case (b)), the accuracy is still ∼10 centimetres. Though this number is based on these image frames that we evaluated, it is safe to say that sub-metre accuracy is achieved with I2-S2 when at least one reference image containing overlapping features with the query is provided. Next, we will show that such reference images can be selected from multiple candidates in a database.



(a) "Office ceiling dataset"



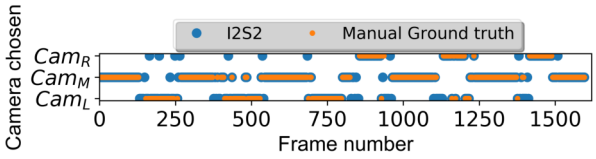
(b) "Mine tunnel multi-traverse dataset"

Figure 9: Sample frames demonstrating the effectiveness of reference select module. In (a) and (b), first row left: Query image from Cam_Q, first row right: Selected reference image. Second row: Query image overlapped with optical flow field w.r.t various reference image candidates in the third row, generated by I2-S2. Third row: Reference image candidates within $\pm t$ frames in the three traverses ($t = 10$ in (a), $t = 3$ in (b), Cam_L: Blue, Cam_M: Green, Cam_R: Red). The image selected is the "current" frame from Cam_R in both cases.

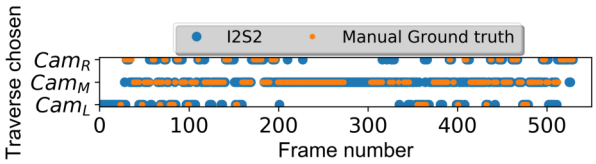
5.2 Reference Select Module Performance

The reference select module was first evaluated using the "office ceiling dataset". As shown in Fig. 9a, the query image was matched to three reference images from each of the three reference traverses and the candidate that offered the most in-lier optical flow vectors was selected. This metric effectively selects the candidate reference that is most similar in terms of viewing angle w.r.t the query image, such as the "current" frame from Cam_R in the example shown in Fig. 9a. Suppose only those reference images from Cam_M were available, in order for the frames from Cam_M to give more consistent optical flow field, the size L_{SR} of the search box would have needed to be dramatically increased to bring the corresponding features (e.g. those on the pipe) into the search range. Now it is unnecessary since a good reference candidate was available from Cam_R. Also note that the mirror image of the pipe reflected by the glass that could cause visual aliasing did not affect I2-S2 in this case.

The reference select module was then applied to the "mine tunnel multi-traverse dataset". An example frame



(a) "Office ceiling dataset"



(b) "Mine tunnel multi-traverse dataset"

Figure 10: Blue dots: Traverse selected for each frame by reference select module. Orange dots: Manually labelled ground truth, if present, means the image from the indicated reference traverse is more similar to the query image than the other two for that frame. No ground truth is indicated when at least two reference images look nearly identical.

for which a best matched reference was correctly identified is shown in Fig. 9b. The results for other frames in these two datasets are summarised in Figures 10a and 10b, respectively. The transition between the references selected are plotted in blue and the manually-labelled ground truth is overlapped in orange. Note that for the "office ceiling dataset", when Cam_Q was placed between Cam_L and Cam_M during the first loop (frames 0 to 799), the selected references were mostly from these two cameras. The selected frames were mostly from Cam_R and Cam_M during the second loop, corresponding to the change in Cam_Q's position. The period of transition in the ground truth is consistent with the period of 130 frames - the artificially generated motion of the crop window (subsection 4.1). It is clear that reference select module correctly captured the "zigzag" motions in both cases. For frames where ground truth is not present, most likely at least two of the reference images are similar. With such frames excluded, the percentage of correct reference selection w.r.t. manually labelled ground truth is 96% on the "office ceiling dataset" and 78% on the "mine tunnel multi-traverse dataset".

The above results show that I2-S2, integrated with the reference select module, constitute a framework that can perform sub-metre underground localisation. Finally, the framework was run on an Intel Core i7-7700K CPU @ 4.20GHz to generate the visual odometry part of the video attachment for this paper, during which the averaged homography output frame rate was ∼0.5 fps. By tweaking the number of sample points (56 in the video) along with sample grid spacing l_G, this frame rate can be adjusted to suit real-time applications. As a comparison, it took MR-FLOW ∼7 minutes to calculate the dense optical flow for a single frame (initialisation included).

6 Discussions

Visual aliasing resulting from repetitive features and occlusions in mine tunnel ceiling images can affect performance of the benchmark method MR-FLOW. Common error messages from MR-FLOW on these images were "too few structure matches" and "too many pixels are occluded". The error in the MR-FLOW optical flow field gets large when the estimated rigidity map is incorrect, and increases with vehicle velocity. I2-S2 is more robust to visual aliasing problems because it does not rely on rigidity estimation or semantics of objects. The matches between image patches are determined based on a holistic impression around the query matches' neighbourhood. This is evident from Figure 7 by comparing the optical flow vectors sampled on objects, such as wires and pipes, calculated by the three methods. Because these wires and pipes are self-similar structures along their own longitudinal directions, the image patch sampled on them are strongly aliased to patches sampled on other parts of the same object. We can see how visual aliasing confused the vanilla method and MR-FLOW by observing the amount of out-lier optical flow vectors that are inconsistent with in-lier vectors as well as the manual ground truth; on the other hand, I2-S2 is more robust and less affected by visual aliasing.

I2-S2 analyses each frame independently and localisation error does not accumulate; it takes almost no time to recover from a mismatched frame, which happens rarely as long as one of the reference candidates overlaps the query image with features less than L_{SR} pixels apart. Such robustness is desirable in automation for continuous mining production. As shown in Fig. 7, erroneous patch matches can be filtered out by RANSAC algorithm and the homography is based solely on high-quality matches.

7 Conclusion

In this paper, I2-S2, a pixel correspondence matching algorithm and an accompanying reference selection module designed for vision-based underground mining vehicle localisation are presented. Robust to visual aliasing, I2-S2 provides localisation refinement down to sub-metre accuracy based on sparse optical flow estimations. No pre-training is required and images are processed individually. The light-weight sequence sampling method is augmented by the coherent-trajectory matching process efficiently implemented with dynamic programming. It is shown to achieve ∼50% more accurate optical flow estimation than the state-of-the-art deep-learning based

MR-FLOW on a mine tunnel dataset while requiring less input information and computation resource. The accuracy of I2-S2 remains stable regardless of flow vector magnitudes. The size of search box can be constrained by incorporating multiple references. The homography estimation w.r.t multiple references can further enhance the localisation accuracy. The uncertainty of vehicle location is insensitive to direction, in contrast to that of a laser scanner, which is often only good in the direction perpendicular to the tunnel.

In future, we will optimise the performance of I2-S2 for a live demonstration inside a mine tunnel. For instance, comparing intra-image sequences captured with different α angles can be a promising pathway to further improvements, especially w.r.t. camera rotation. It also becomes possible to use the optical flow information from I2-S2 as a visual odometry prior, useful for integration with other localisation modules for a fully automated underground mining system.

References

[Bay *et al.*, 2006] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Eur. Conf. Comput. Vis.*, pages 404–417. Springer, 2006.

[Birchfield, 2007] Stan Birchfield. KLT: An implementation of the Kanade-Lucas-Tomasi feature tracker. *http//www. ces. clemson. edu/˜ stb/klt/*, 2007.

[Caruso *et al.*, 2015] D Caruso, J Engel, and D Cremers. Large-Scale Direct SLAM for Omnidirectional Cameras. In *Int. Conf. Intell. Robot. Syst.*, 2015.

[Cummins and Newman, 2008] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *Int. J. Rob. Res.*, 27(6):647–665, 2008.

[Daoust *et al.*, 2016] Tyler Daoust, François Pomerleau, and Timothy D Barfoot. Light at the End of the Tunnel: High-Speed Lidar-Based Train Localization in Challenging Underground Environments. In *Comput. Robot Vis. (CRV), 2016 13th Conf.*, pages 93–100. IEEE, 2016.

[Dosovitskiy *et al.*, 2015] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2758–2766, 2015.

[Jacobson *et al.*, 2018] Adam Jacobson, Fan Zeng, David Smith, Nigel Boswell, Thierry Peynot, and Michael Milford. Semi-Supervised SLAM: Leveraging Low-Cost Sensors on Underground Autonomous Vehicles for Position Tracking. In *Int. Conf. Intell. Robot. Syst.*, page In Press, Madrid, Spain, 2018.

[Lavigne and Marshall, 2012] N. James Lavigne and Joshua A. Marshall. A landmark-bounded method for large-scale underground mine mapping. *J. F. Robot.*, 29(6):861–879, 2012.

[Lucas *et al.*, 1981] Bruce D Lucas, Takeo Kanade, and Others. An iterative image registration technique with an application to stereo vision. 1981.

[Milford and Wyeth, 2012] Michael J. Milford and Gordon F. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *Proc. - IEEE Int. Conf. Robot. Autom.*, 2012.

[Moridi *et al.*, 2015] Mohammad Ali Moridi, Youhei Kawamura, Mostafa Sharifzadeh, Emmanuel Knox Chanda, Markus Wagner, Hyongdoo Jang, and Hirokazu Okawa. Development of underground mine monitoring and communication system integrated ZigBee and GIS. *Int. J. Min. Sci. Technol.*, 25(5):811–818, 2015.

[Muja and Lowe, 2014] Marius Muja and David G Lowe. Scalable Nearest Neighbor Algorithms for High Dimensional Data. *Pattern Anal. Mach. Intell. IEEE Trans.*, 36, 2014.

[Mur-Artal Raúl and Tardós, 2015] Montiel J M M Mur-Artal Raúl and Juan D Tardós. {ORB-SLAM}: a Versatile and Accurate Monocular {SLAM} System. *IEEE Trans. Robot.*, 31(5):1147–1163, 2015.

[Sergeant *et al.*, 2016] James Sergeant, Gary Doran, David Thompson, Abigail Allwood, Christopher Lehnert, Ben Upcroft, and Michael Milford. Towards multimodal and condition-invariant vision-based registration for robot positioning on changing surfaces. *Australas. Conf. Robot. Autom. ACRA*, 2016.

[Wulff *et al.*, 2017] Jonas Wulff, Laura Sevilla-Lara, and Michael J Black. Optical Flow in Mostly Rigid Scenes. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, jul 2017.

[Zeng *et al.*, 2017] Fan Zeng, Adam Jacobson, David Smith, Nigel Boswell, Thierry Peynot, and Michael Milford. Enhancing Underground Visual Place Recognition with Shannon Entropy Saliency. In *Australas. Conf. Robot. Autom. ACRA*, Sydney, Australia, 2017.

[Zhang and Singh, 2015] Ji Zhang and Sanjiv Singh. Visual-lidar odometry and mapping: Low-drift, robust, and fast. *Proc. - IEEE Int. Conf. Robot. Autom.*, 2015-June(June):2174–2181, 2015.

[Zlot and Bosse, 2014] Robert Zlot and Michael Bosse. Efficient large-scale three-dimensional mobile mapping for underground mines. *J. F. Robot.*, 31(5):731–752, 2014.