Enhancing Underground Visual Place Recognition with Shannon Entropy Saliency

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

Autonomous vehicles are increasingly being used in the underground mining industry, but competition and a challenging market is placing pressure for further improvements in autonomous vehicle technology with respect to cost, infrastructure requirements, robustness in varied environments and versatility. In this paper we introduce several new vision-based techniques for visual place recognition on underground mining vehicles that improve on current available technologies while only requiring camera input. We present a Shannon Entropy-based salience generation approach that enhances the performance of single image-based place recognition by selectively processing image regions. We then add a learning-based scheme realised by Support Vector Machines (SVMs) that filters problematic images from both the reference map databases and from live processing. We evaluate the approaches on new large real-world underground vehicle mining datasets, demonstrating ∼107% improvement in area under the precision-recall curve than a state-of-the-art place recognition algorithm, and demonstrate its generality on a non-mining-based benchmark dataset. Together this research serves as a step forward in developing domain-appropriate improvements to existing state-of-the-art place recognition algorithms that will hopefully lead to improved efficiencies in the mining industry.

*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.

1 Introduction

The mining industry is one of the major contributors to the Australian economy, with underground mining playing an increasingly significant part. After its initial construction, the daily operation of an underground mine site typically involves workers driving heavy vehicles such as load-haul-dump (LHD) loaders inside underground mine tunnels (Fig. 1a) to perform various manoeuvres such as digging (Fig. 1b) and transporting. Automating some or all of these tasks can reduce human labour and error, making the mine site safer and more productive. One of the keys to complete automation is reliable localisation of the mining vehicle in the underground environment, where a commonly-adopted source of localisation information - the Global Positioning System (GPS) is not available. Existing underground localisation technologies mostly depend on either Received Signal Strength Indicator (RSSI) [Liu *et al.*, 2010; Rohrig and Muller, 2009] or Time of Arrival (ToA) based [Sathyan *et al.*, 2011] trilateration systems. The widely deployed state-of-the-art, including Wireless Sensor Networks (WSNs) [Moridi *et al.*, 2015; Li and Liu, 2009] and Radio-Frequency IDentifications (RFIDs) [Rusu and Hayes, 2011; Kent, 2011; Zhou and Shi, 2009], can be less reliable due to interference from the rock mass, and typically require substantial initial investment in time and infrastructure, e.g., setting up mesh of beacons such as Wireless Local Area Network (WLAN) Access Points (APs), and characterizing wireless propagation parameters such as pathloss [Forooshani *et al.*, 2013]. The initial cost grows with the size of the mine site, eventually making these localisation systems difficult and expensive to maintain. In addition, for better localisation performance, the beacon density needs to be increased accordingly, as shown in the RFID-based underground mine localisation field test in [Lavigne and Marshall, 2012]. A low-cost plug-and-play localisation technology that dynamically builds up the map without asking for help from human operators is more desirable. Like in other domains, vision sensors such as cameras are at-

tractive because they provide rich information with high replaceability, small foot-print and minimal disturbance or interference to the mine site.

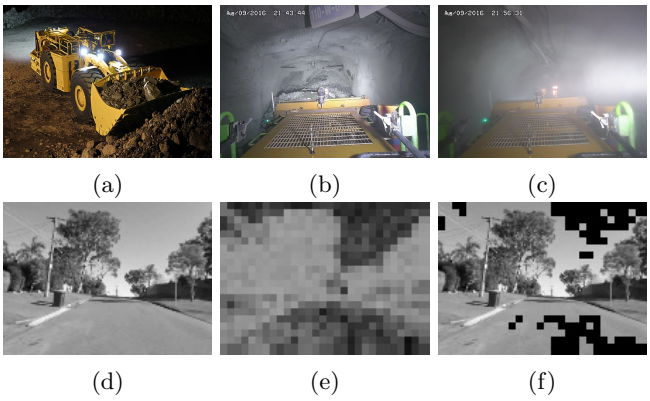


Figure 1: (a) A photo of a load-haul-dump (LHD) loader on underground mine site. (Source: Caterpillar.) (b) A clear photo taken by camera mounted on an LHD loader. (c) Photo taken when LHD loader is travelling through dust in a tunnel. (d-f) SESAME algorithm removes inconspicuous image patches from query image (d) according to information entropy metric (e), like sprinkled sesames (f).

Underground mine tunnels offer unique challenges for visual place recognition, however. First, unlike street or indoor environments where abundant feature-rich landmarks provide useful cues for localisation, the lack of feature inside the tunnels creates perceptual aliasing that dramatically reduces the reliability of feature-based visual place recognition systems. Second, inevitable dust (Fig. 1c) generation not only blocks useful landmarks but also forms transient false features that compromise place recognition accuracy. Last but not least, the camera can occasionally be overwhelmed with partial or complete blockage, interference such as head-lights from another vehicle, sparks generated during the mining process and so on. We start to tackle these challenges by noting that some image pixels are useful because they correspond to salient landmarks. Conversely, some pixels contain little useful information (Figs. 1d to 1e). Pixels that correspond to transient features can even contribute negatively. In the research we show that removing such pixels (Fig. 1f) not only saves computation resources, but also improves the performance of the underground place recognition system. On a mine site dataset, ∼8.8% improvement over the holistic single-image-comparison approach has been observed. The generality of this technique has also been demonstrated with a ∼40% improvement on a completely different dataset.

The paper proceeds as follows. Section II reviews previous works on confined space localisation and relevant literature on saliency-based computer vision algorithms. Section III elaborates the detailed implementations of various components in our approach. Section IV gives the configurations of experimental set-up to verify our approach on a mine site and a benchmark dataset, with the results of experiments presented in Section V. A brief discussion on the results and future work can be found in Section VI.

2 Literature Review

Extensive research has focused on visual mapping and localisation in confined spaces such as localisation of in-pipe [Lee *et al.*, 2011; Paul *et al.*, 2010], surgery [Taylor *et al.*, 2016; Wang *et al.*, 2017] and mining robots. For the in-pipe robot design in [Lee *et al.*, 2011], localisation has been accomplished using landmarks such as elbows and branches in the pipe combined with orientation information from an IMU (Inertia Measurement Unit) sensor. Laser scanning sensors have been used in [Paul *et al.*, 2010] to construct an accurate map of the bridge structure for localisation of a bridge inspection robot. It took 30 to 50 repetitions of scanning to produce a map that is accurate enough for the application. Detailed scannings like this are less feasible on the scale of underground mines. A laparoscopic visual tracking system has been designed in [Wang *et al.*, 2017] to "localise" the surgical instrument within its work space. A special marker was embedded to provide computer vision features that were extracted in Hue-Saturation-Value (HSV) space for illumination invariance. However, it is less feasible to implement this marker approach in the underground mining context.

A hierarchical system called "Opportunistic Localization" has been built in [Duff *et al.*, 2003] to localise mining vehicles based on wheel-odometry integrated with prediction of possible forward routes from the current location belief. It was assumed that localisation accuracy is more crucial at intersections than within a tunnel. This assumption may not be valid in circumstances where prevention of collision between pedestrian workers and the vehicle is required. [Gadd and Newman, 2015] presents a texture-tracking visual teach and repeat (VT&R) system for vehicles travelling in fixed industrial environment; typical underground mine sites are noisier and less controlled. SLAM++ [Salas-Moreno *et al.*, 2013] tracks salient objects in the scene to provide crucial localisation cues, but underground mines typically do not present easy-to-track objects. The classic feature-based place-recognition algorithm FAB-MAP [Cummins and Newman, 2008] detects the existence of "visual words", discarding geometric relative location information of features that are crucial in underground environment - for example, the same "visual words" may appear in the field of view for a range of contiguous lo-

cations in a long tunnel. Localisation in dark environments through extracting larger-scale features has been demonstrated in [Nelson *et al.*, 2015] and [Li *et al.*, 2015], but their target applications were not general enough for the techniques to be directly transferable to underground mine sites.

There are previous works on applying saliency criteria to image processing in other applications. In [Pepperell *et al.*, 2014], pixels corresponding to sky were removed before further processing, but it was done using a dedicated technique that specifically tackled the skies. Multi-scaled saliency masks have been shown to work on object detection [Frintrop *et al.*, 2014], a similar idea is incorporated in our approach (Section 3.3). In [John *et al.*, 2015], salient features such as traffic lights were extracted using Convolutional Neural Network (CNN) for localisation on urban streets. In [Milford *et al.*, 2014], a patch verification method was used in conjunction with "edge" and "human" saliency masks, a similar idea is adopted by our approach in the underground mining context. Details of our saliency-based patch filtering approach is described next.

3 Approach

In this section, the holistic single-image-comparison based place recognition approach will be described first, followed by the concept and components of SESAME - *S*hannon *E*ntropy *SA*liency *ME*thod (Figs. 1d-1f), complemented by an SVM classifier trained to help maintain the quality of images in the database. SESAME on an alternative deep learning technique is lastly discussed.

3.1 The Holistic Single-Image-Comparison Approach

In the holistic single-image-comparison approach (Fig. 2 with yellow boxes omitted), direct single image comparison (without incorporation of temporal information) is used to map a query image j to its global best match $l(j) = i_{best}$ in the database (a dictionary containing images labelled with their ground truth location), in terms of lowest Sum of Absolute Difference (SAD) score $S(i, j)$ [Milford and Wyeth, 2012]. To tolerate some amount of viewing angle change, the images being compared are shifted w.r.t each other for offset values within a search window σ [Milford, 2013], with the minimum SAD score stored in a confusion matrix D. Apart from being cropped and resized to $R_x \times R_y$ pixels, query and database images are all converted to gray-scale and patch-normalised in the preprocessor to enhance robustness w.r.t lighting and other changes in environment before the intensity of every pixel is compared unbiasedly. This holistic image comparison with patch normlisation has been proposed in the SeqSLAM algorithm [Milford and Wyeth, 2012]. Direct comparison of whole images retains as much as possible the relative positional information of objects so that perceptual aliasing can be reduced. However, in contrast to SeqSLAM, single image comparison without temporal information is used in this work so that the effect of patch saliency can be emphasized.

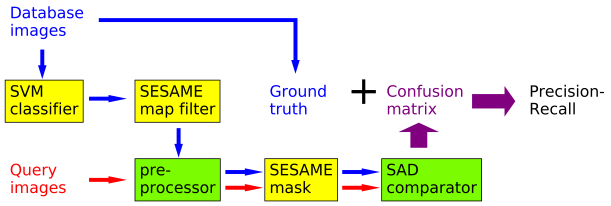


Figure 2: Schematic drawing of the proposed visual place recognition approach. Intermediate result corresponding to database and query images are shown in blue and red, respectively. Modules of the holistic single-image-comparison approach are in green boxes, and the newly proposed SESAME modules are in yellow boxes.

In the context of underground visual place recognition, SAD score $S(i, j)$ are often found to consist of contributions from two kinds of intensity difference between query patch p_{qr} and database patch p_{db} (1): Salient patches $\{p_s\}$ that are key for localisation (2) and textureless patches $\{p_{\bar{s}}\}$ that contain little useful information (3). Including the latter into SAD only adds noise to the process. Intuitively, it is better to assign little or no weight in the composition of SAD scores for inconspicuous patches in $\{p_{\bar{s}}\}$ and only account for the SAD contributed by pixels in salient patches $\{p_s\}$, as shown in (4).

$$l(j) = \arg\min_i S(i, j) = \arg\min_i [S_s(i, j) + S_{\bar{s}}(i, j)] \quad (1)$$

$$S_s = \sum_{p_{db}, p_{qr} \in \{p_s\}} (\sum\sum |p_{qr} - p_{db}|) \quad (2)$$

$$S_{\bar{s}} = \sum_{p_{db} \text{ or } p_{qr} \notin \{p_s\}} (\sum\sum |p_{qr} - p_{db}|) \quad (3)$$

$$l(j) = \arg\min_i S_s(i, j). \quad (4)$$

For instance, one would rely more on the location labels written on the overhanging signs (Fig. 3a) in an underground mine tunnel, instead of uniform-coloured patches of walls on the two sides of the vehicle. Furthermore, objects like the overhanging wirings (Fig. 3a) could also be salient landmarks, just like numbers on the signs. Computers are more capable to recognize, remember and process them than human eyes.

3.2 The Shannon Entropy SAliency MEthod (SESAME)

Our proposed method of accentuating salient objects over monotonic image regions is to filter image patches according to their Shannon Entropy $H(p)$: After preprocessing steps such as gray-scale conversion and resizing, a query image is divided into small individual patches for which Shannon Entropy is calculated using (5):

$$H(p) = -\sum_{i=0}^{255} P_p(b_i) log P_p(b_i), \tag{5}$$

in which b_i is one of the 256 possible pixel intensity value bins for an 8-bit monocolour image, $P_p(b_i)$ = [number of pixels with intensity = b_i]/$L_{en}(p)^2$ is the probability for a pixel to take on such intensity value within the patch p, and $L_{en}(p)$ is the side length of patch p that is used for entropy calculation, not to be confused with the patch size used in patch normalisation $L_{norm}(p)$. Patch size will refer to the former, if not indicated otherwise in the remainder of this text. A visualisation of the patch-wise entropy is shown in Figure 3b. The rest of the process, including patch normalisation and viewing angle shift compensation, follows the holistic single-image-comparison approach, except that the patches with lower Shannon Entropy value are discarded when calculating the SAD score. The percentage of discarded patches is controlled by a data-independent parameter $r_{threshold} \in (0, 1)$, which is the relative threshold to accept a patch as salient, based on its Shannon Entropy value w.r.t the patches with maximum and minimum Shannon Entropy among all patches in the entire image. Relative instead of absolute threshold is used so as to discard image patches based on quality instead quantity. There are other flavors of implementation such as setting an absolute Shannon Entropy threshold that may be helpful in some cases and worth further investigation, though in those cases the value of optimum absolute threshold will have more dependency on specific application details, such as data quality, image colour depth, etc.

A binary salient mask M (Fig. 3c) is used to indicate the locations of salient pixels that should be accounted in SAD calculation. The image in Figure 3a with inconspicuous patches removed is shown in Figure 3d. Since the amount of pixels removed could vary for different images, the SAD value should be normalised by the number of salient pixels that have their contribution, as shown in (6).

$$S = \min_{\Delta x, \Delta y \in \sigma} (I_{qr}(y + \Delta y, x + \Delta x) - I_{db}(y, x)) * (M(y + \Delta y, x + \Delta x)) / \sum_{i,j} M(i, j). \tag{6}$$

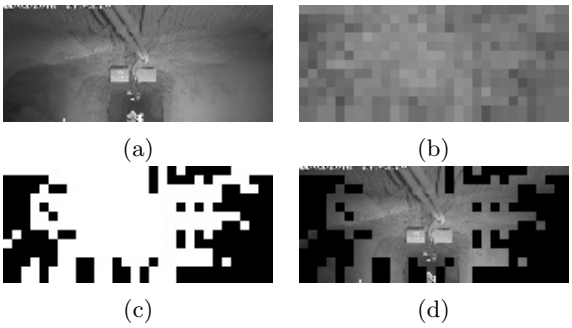


Figure 3: (a) An image taken in an underground tunnel showing salient features such as overhead signs and wirings. (b) A visualisation of the entropy value for the image in (a) with $L_{en}(p) = 5$ pixels, lighter color means larger entropy. (c) The binary bask M generated using (b) and $r_{threshold}$ of 0.5. (d) The processed image with lower saliency patches removed, to be used for SAD comparison.

3.3 Hierarchical SESAME

The patch size $L_{en}(p)$ plays an important role in deciding which pixels are discarded. Since salient features show up in different sizes, using a pre-defined patch size $L_{en}(p)$ limits the flexibility of the approach. For instance, if $L_{en}(p)$ is too small, saturated pixels inside large salient features which could potentially increase matching accuracy could be removed. On the other hand, if $L_{en}(p)$ is increased to enclose the larger salient features, some salient features that happen to lie close to a plain patch could be implicated. Therefore, it is better to implement a multi-level hierarchical masking process to only remove pixels that are plain regardless of scale, as shown in (7), with n being the total number of saliency masks to be used. Figure 4 shows the retaining of low entropy patches that are salient on a larger scale with Hierarchical SESAME.

$$M = M_{L_{en}(p)_1} \cap M_{L_{en}(p)_2} \cap ... \cap M_{L_{en}(p)_n} \tag{7}$$

3.4 Saliency of Database Images

Apart from selecting an appropriate method for image matching, the saliency of database images also plays an important role. Problematic images sometimes can attract a lot of erroneous matches. For example, a mostly dark or bright database image caused by typical disturbances on mine sites such as obstruction, dust, lens glare, head-light from other vehicles, etc., will sometimes be the global best match for a large amount of queries (i.e., the corresponding row in the confusion matrix D exhibit a low SAD score relative to all other places), this is especially harmful to Precision-recall if it happens to be a best match for an extended stationary part of the query

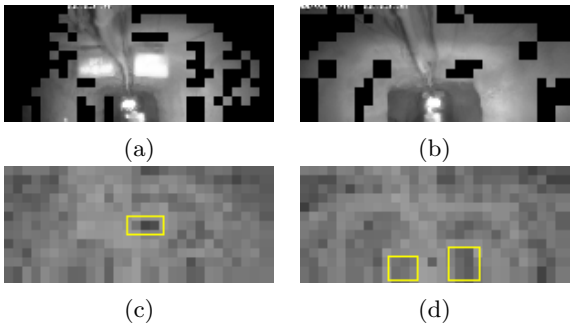


Figure 4: (a-b) Images filtered by hierarchical SESAME. (c) Entropy visualisation for (a), the useful low entropy patches corresponding to overexposed sign enclosed by yellow box is retained after filtering. (d) Entropy visualisation for (b), the useful low entropy patches corresponding to underexposed tunnel enclosed by yellow boxes are retained after filtering.

set when the vehicle is digging. These problematic images should be excluded from the database. Problematic images in the query only cause trouble during the image-matching process of themselves in our single-image based approach, therefore no extra step was taken to exclude such images from the query.

Since the same place is usually repeatedly visited in the dataset due to the nature of mining process, there could be multiple images that correspond to the same view of the same location. It is therefore possible to choose among those images and only keep the most representative in the database for each unique view, minimising the number of problematic images. The details of SESAME map filter module are shown in Algorithm 1, in which A_{th} is the Adjacency threshold (images will be considered as belonging to the same place if distance between ground truth labels is less than A_{th}), V_{th} is the Visual Similarity threshold (images will be considered as belonging to the same view if SAD is less than V_{th}), and S_{th} is the Saliency threshold of the Saliency Score SS (8), in percentage of its maximum possible value SS_m (9).

$$SS = \sum_{i,j} M(i,j) \tag{8}$$

$$SS_m = 255 * R_x * R_y \tag{9}$$

As a result, SESAME can automatically exclude those pure-coloured images from database since they have low Shannon Entropy (Fig. 5a and 5b). Even in cases when such images are not removed from database, part of the problematic portion can be removed by SESAME (Fig. 5c and 5d).

Input: $inImages, A_{th}, V_{th}, S_{th}$
Output: $uniqueImages$
$uniqueImages \leftarrow \emptyset$
for $inImage \in inImages$ **do**
  $novelView \leftarrow true$
  for $image \in uniqueImages$ **do**
    if $distance(inImage, image) < A_{th}$ **and** $similarImg(image, inImage, V_{th})$ **then**
      $novelView \leftarrow false$
      if $inImage.saliency > image.saliency$ **then**
        $image \leftarrow inImage$
        break
  if $novelView$ **and** $inImage.saliency > S_{th}$ **then**
    $uniqueImages \leftarrow uniqueImages \cup inImage$

Algorithm 1: *BuildSalientDatabase*

3.5 Thorough Removal of Problematic Images

However, images such as those shown in (Fig. 6a) are particularly tricky. Due to the gradual intensity change, the Shannon Entropy is actually not low enough to exclude them. Although some information exist in the histogram equalised version of the image (Fig. 6b), it makes negligible helpful contribution to the SAD comparison after patch normalisation. Thus in addition to SESAME, a linear SVM is trained as an image classifier to further remove such problematic images from the database.

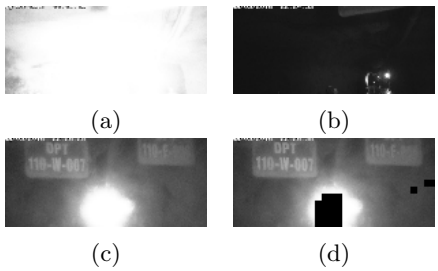


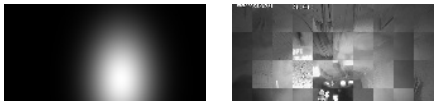
Figure 5: Problematic database images discarded by SESAME. (a) low saliency and too bright, (b) low saliency and too dark. (c) A problematic image with glare. (d) Hierarchical SESAME removes patches of glare in (c) and keeps salient features.

The training of the SVM starts with the generation of data that include training and testing sets of "good" and "bad" images. The "bad" training samples (Fig. 7a) are generated by a Gaussian blob with random standard deviation at random locations on the image, simulating the saturation effect of bright light shining on the camera.

(a) (b)

Figure 6: (a) Example of a problematic image that should be excluded from the database. The bright light is not a fingerprint of the place but attracts lots of false matches. (b) A histogram-equalised version of (a).

Their "good" counter-parts (Fig. 7b) are generated by stitching together random-sized patches from randomly selected source images that represent typical scenes in the tunnel. In our case, 20 such images are used as shown in Figure 8, more could be used for performance enhancement. The patches are sampled from the neighborhood



(a) (b)

Figure 7: Examples of artificially generated (a) "bad" and (b) "good" images used to train the SVM image classifier.

of its corresponding location in the selected source image. In this way, large quantities of training data can be obtained without the need for collection of real samples and tedious manual labelling work.

3.6 SESAME on Deep Learning Neural Network

A deep learning neural network (hereafter Deep Net) [Sünderhauf *et al.*, 2015] can help trade-off computation resources for better results. When lower speed and large memory is allowed, Deep Net may be preferred. Hence the effectiveness of SESAME on Deep Net (that replaces the "SAD comparator" in Fig. 2) is also evaluated in the experiments described next.

4 Experimental Setup

The effectiveness of SESAME on both holistic single-image-comparison and Deep Net approaches was evaluated by running localisation experiments on an Underground Mine dataset. For Deep Net, the "fc6" layer in a variant of the Deep Net in [Sünderhauf *et al.*, 2015] was used to produce the confusion matrix for Precision-recall curve generation. The *general* effectiveness of SESAME was evaluated with a benchmark dataset featuring a completely different domain. The details of experiments are covered in this section.

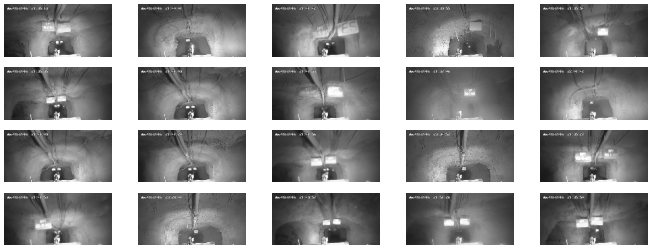


Figure 8: Manually selected images that are used during artificial generation of "good" images for SVM training.

4.1 Datasets

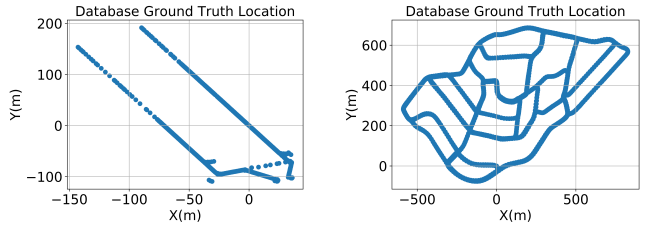
The Underground Mine Dataset The underground Mine Dataset photos were taken by cameras mounted on the back of an LHD loader during its daily operation. The dataset contains a 104-minute-long footage of ∼15.6 kilometers of back and forth travel inside 2 long tunnels (Fig. 9a) and a small intersection region where the LHD loader dumped the load. The 20,000 photos (size: 1280×960 pixels) were sub-sampled (step length = 5) for database construction, and a non-overlapping (offset = 2) part were sub-sampled (step length = 5) as query images. Note in this experiment, no temporal correlation is drawn upon and each image is treated independently. Therefore it is not necessary to sequentially process query images to get the same result.

The Benchmark Dataset The images of the "St Lucia Multiple Times of Day" dataset [Glover *et al.*, 2010] were taken by a front-facing camera mounted on top of the car during a ∼17.6 kilometer-travel in Brisbane, Australia (Fig. 9b). The captured video for each traverse was converted to ∼20,000 images. Two traverses - 1009091410, 2108091210 ("DD/MM/YY 24HOUR") from the dataset were sub-sampled (step length = 5) for database construction and another traverse (1109091545) was sub-sampled (step length = 10) as the query.

4.2 Ground Truth

The Underground Mine Dataset Ground truth locations of the Underground Mine dataset is measured using an accurately surveyed beacon-infrastructure based trilateration localization system.

The Benchmark Dataset The planar coordinates were approximated by scaling the processed GPS data that come along with the "St Lucia Multiple Times of Day" dataset.



(a) (b)

Figure 9: Map of database ground truth location for (a) the Underground Mine dataset and (b) the "St Lucia Multiple Times of Day" dataset.

4.3 SVM Classifier Training

The training took 20 epochs of Stochastic Gradient Descent (SGD) to converge, with mini-batch size of 20 and variable learning rate between 0.0001 to 0.01. The training continued until a 95% accuracy was reached for a separate testing set, after which the performance of the SVM was evaluated using an evaluation set of 220 images including artificially generated ones in both "good" and "bad" categories that the classifier had not seen before, mixed with the 20 ("good") images that were used to generate the "good" training samples (Fig. 8). The trained SVM was then used to filter candidate database images. The SVM classifier was not used on "St Lucia Multiple Times of Day" dataset because it contains few problematic images.

4.4 Image Pre-Processing and Comparison

In our current implementation, saliency masks M_5, M_{10}, M_{20} with $L_{en}(p) = 5, 10, 20$ are calculated, and the final saliency mask M is the intersection of the three (10). In other words, a pixel will be removed only when it lies simultaneously in plain patches of sizes 5, 10, and 20 pixels.

$$M = M_5 \cap M_{10} \cap M_{20} \tag{10}$$

4.5 Comparison with FABMAP

In order to benchmark the performance of SESAME with FAB-MAP [Cummins and Newman, 2008], the same set of raw unfiltered camera images for both database and query were converted to grayscale, cropped in exactly the same way as SESAME, resized to 640×304 pixels and converted to videos. OpenFABMAP [Glover *et al.*, 2012] was then run with semantic vocabulary generated with these input videos for its best performance.

4.6 Parameters

The parameters used in the current SESAME implementation for the Underground Mine dataset are summarised in Table 1. These parameters are heuristically selected but have not been optimized, neither have they been fine-tuned to work best in the particular dataset. For instance, R_x and R_y are chosen to reduce the image size and speed up processing, while keeping the aspect ratio unchanged; the adjacency threshold is chosen so that different places in the resulting database exhibit visible difference. The selection of maximum tolerance for True Positives has taken into account the length of travel along which similar views can be observed inside the tunnel. The results obtained using these parameters are presented next in Section 5.

Table 1: PARAMETER LIST

Parameter	Value	Unit	Description
R_x	160	pixel	Reduced image width
R_y	76	pixel	Reduced image height
$L_{norm}(p)$	5	pixel	patch side length used in patch normalisation
A_{th}	0.5	meter	Adjacency threshold
V_{th}	0.25		Visual Similarity threshold
S_{th}	50%		Saliency threshold
T_{tp}	15	meter	Max. tolerance for True Positives

5 Results

In this section, the results of the experiments are presented, including the performance of the SVM classifier and a comparison of Precision-recall curves achieved by various methods described in the previous section.

5.1 SVM Classification Effectiveness

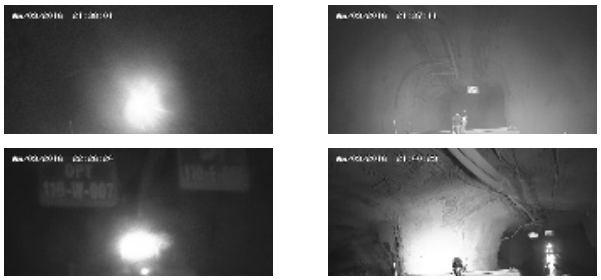
The accuracy of the trained SVM is first evaluated. Among all images (124 out of 220) classified by the trained SVM to be "good", 115 (93%) are actually "good"; Among all the images (85 out of 220) classified to be "bad", 84 (99%) are actually "bad". 11 out 220 images are found to be ambiguous by the SVM, including one of the 20 images used as "good" image patch source (the remaining 19 are all correctly classified). When challenged with real-world images, the SVM could correctly pick out "bad" images such as those shown in Figure 10a, along with some misclassified "good" images (Fig. 10b), which is not a big problem since there are adequate amount of images to cover the entire map.

5.2 Computational Resource Usage

The apparent computation resource usages by typical runs of three groups of methods in our experiments have been recorded and shown in Figure 11.

5.3 Precision-Recall Curves

The Precision-recall curves of the holistic single-image-comparison approach, Deep Net (both with and w/o various levels of SESAME), and OpenFABMAP as reference, are plotted in Figure 12 for the Underground Mine dataset. Results for the benchmark dataset are in Figure 14. The areas under these Precision-recall curves are estimated and shown in Figures 13 and 15, respectively. Part of these localisation results can be visualized in the video companying this paper.



(a) (b)

Figure 10: SVM map image classification result. (a) Correctly discarded images. (b) Mistakenly discarded images.

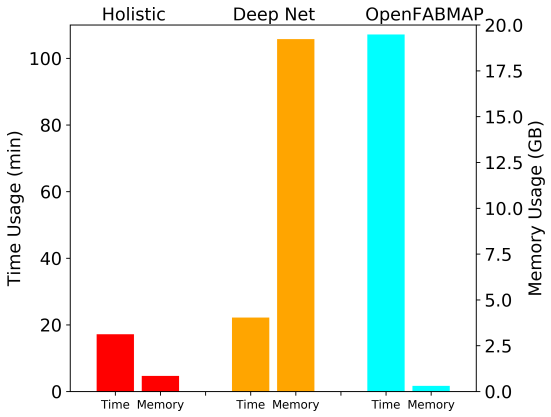


Figure 11: Comparison of computation resource consumption in our experiments by three groups of methods.

6 Discussion and Future Work

The results from the previous section demonstrated that, as a suitable approach for applications that require real-time output and small memory consumption (such as underground mining), the holistic single-image-comparison method with a light to moderate $r_{threshold} \in [0.1, 0.5]$ produces desirable results in terms of area under the Precision-recall curve; while higher levels of SESAME, as expected, converge to the extremity of all patches being

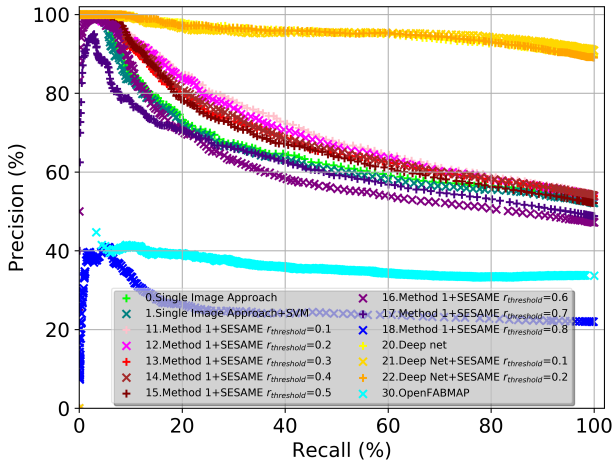


Figure 12: Precision-recall curves comparing performance of Method 0: holistic single-image-comparison approach (without SVM classifier), Method 1: Method 0 with SVM classifier, Methods 11-18: Method 1 with various $r_{threshold}$ levels of SESAME (with SVM classifier), Method 20: Deep Net, Methods 21-22: Deep Net with SESAME, and Method 30: OpenFABMAP, for the Underground Mine dataset.

discarded. A lower level of SESAME (relative threshold $r_{threshold} \in [0.1, 0.2]$) is also effective on Deep Net, which would be preferable for applications that allow slower output rate and large memory usage. Both methods with SESAME enhancement significantly out-performed OpenFABMAP. Results show that SESAME, assisted by

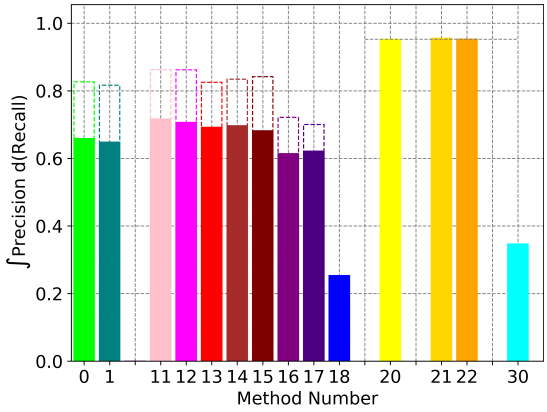


Figure 13: Solid bars: Area under the Precision-recall curves in Figure 12. Hollow dashed regions: Further improvement when temporal information is used.

an SVM database image classifier, is an effective add-on to both holistic single-image-comparison and Deep Net approaches for accurate and reliable localisation in challenging underground environment. The generality of Shannon Entropy-based saliency computation tech-

nique is further demonstrated with the "St Lucia Multiple Times of Day" dataset, with $r_{threshold} \in [0.5, 0.7]$ producing better results. Although a camera-based system is used in the demonstration, there is no apparent barrier for SESAME to be applied on images obtained with laser-based systems such as those in [McManus *et al.*, 2013]. The lightweight SVM classifier used in this work can sometimes lower performance if no "bad" images are present in the database and "good" images are taken out by mistake, this can be improved in the future by replacing the SVM with a more sophisticated neural network. Additionally, temporal information that has not been utilised in this work can be incorporated to eliminate erroneous jumps between far-away locations that are impossible given the vehicle speed, from the set of localisation results. A significant performance boost has already been achieved with a primitive implementation of the said temporal filter, shown as hollow dashed bars in Figure 13. Sensor fusion with IMUs, Lidar and other sources of information integrated with probabilistic filtering also promise potential further improvements when combined with SESAME.

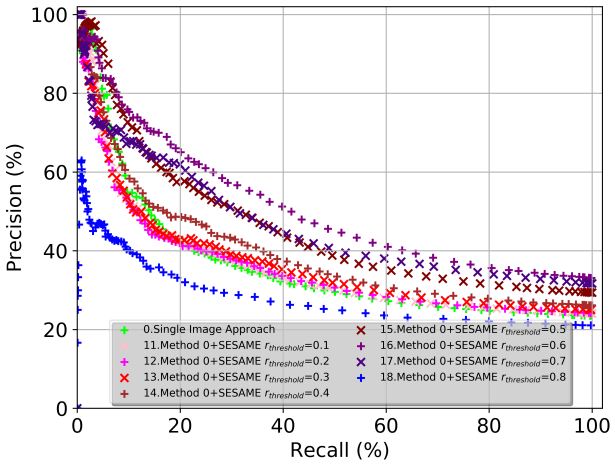


Figure 14: Precision-recall curves comparing performance of Method 0: Holistic single-image-comparison approach, and Methods 11-18: Method 0 with various $r_{threshold}$ levels of SESAME, for the "St Lucia Multiple Times of Day" dataset.

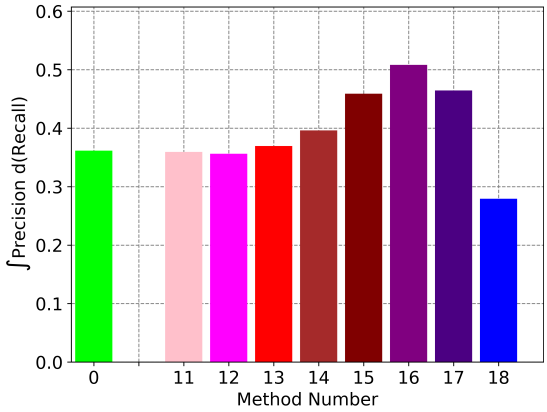


Figure 15: Area under the Precision-recall curves in Figure 14.

Acknowledgments

The authors would like to thank Advance Queensland Innovation Partnerships grant from the Queensland Government, Mining3, Caterpillar, ARC Future Fellowship, and the Queensland University of Technology.

References

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

[Duff *et al.*, 2003] E.S. Duff, J.M. Roberts, and P.I. Corke. Automation of an underground mining vehicle using reactive navigation and opportunistic localization. *IEEE Int. Conf. Intell. Robot. Syst.*, 4(July 1998):3775–3780, 2003.

[Forooshani *et al.*, 2013] Arghavan Emami Forooshani, Shahzad Bashir, David G. Michelson, and Sima Noghanian. A survey of wireless communications and propagation modeling in underground mines. *IEEE Commun. Surv. Tutorials*, 15(4):1524–1545, 2013.

[Frintrop *et al.*, 2014] Simone Frintrop, Gm Garca, and Ab Cremers. A Cognitive Approach for Object Discovery. In *IEEE Int. Conf. Pattern Recognit.*, pages 2329–2334, 2014.

[Gadd and Newman, 2015] Matthew Gadd and Paul Newman. A framework for infrastructure-free warehouse navigation. *Proc. - IEEE Int. Conf. Robot. Autom.*, 2015-June(June):3271–3278, 2015.

[Glover *et al.*, 2010] Arren J. Glover, William P. Maddern, Michael J. Milford, and Gordon F. Wyeth. FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day. In *Proc. - IEEE Int. Conf. Robot. Autom.*, 2010.

[Glover *et al.*, 2012] Arren Glover, William Maddern, Michael Warren, Stephanie Reid, Michael Milford, and Gordon Wyeth. OpenFABMAP: An open source toolbox for appearance-based loop closure detection. *Proc. - IEEE Int. Conf. Robot. Autom.*, pages 4730–4735, 2012.

[John *et al.*, 2015] Vijay John, Keisuke Yoneda, Zheng Liu, and Seiichi Mita. Saliency Map Generation by the Convolutional Neural Network for Real-Time Traf-

fic Light Detection Using Template Matching. *IEEE Trans. Comput. Imaging*, 1(3):159–173, 2015.

[Kent, 2011] Denis Kent. Digital networks and applications in underground coal mines. *11th Undergr. Coal Oper. Conf.*, pages 181–188, 2011.

[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.

[Lee *et al.*, 2011] Dong Hyuk Lee, Hyungpil Moon, and Hyouk Ryeol Choi. Autonomous navigation of in-pipe working robot in unknown pipeline Environment. *Proc. - IEEE Int. Conf. Robot. Autom.*, pages 1559–1564, 2011.

[Li and Liu, 2009] Mo Li and Yunhao Liu. Underground coal mine monitoring with wireless sensor networks. *ACM Trans. Sens. Networks*, 5(2):1–29, 2009.

[Li *et al.*, 2015] Jie Li, Ryan M Eustice, and Matthew Johnson-roberson. High-Level Visual Features for Underwater Place Recognition. *Int. Conf. Robot. Autom.*, pages 3652–3659, 2015.

[Liu *et al.*, 2010] Zhigao Liu, Chunwen Li, Danchen Wu, Wenhan Dai, Shaobo Geng, and Qingqing Ding. A wireless sensor network based personnel positioning scheme in coal mines with blind areas. *Sensors (Switzerland)*, 10(11):9891–9918, 2010.

[McManus *et al.*, 2013] Colin McManus, Paul Furgale, and Timothy D. Barfoot. Towards lighting-invariant visual navigation: An appearance-based approach using scanning laser-rangefinders. *Rob. Auton. Syst.*, 61(8):836–852, 2013.

[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.

[Milford *et al.*, 2014] Michael Milford, Eleonora Vig, Walter Scheirer, and David Cox. Vision-based simultaneous localization and mapping in changing outdoor environments. *J. F. Robot.*, 31(5):814–836, 2014.

[Milford, 2013] M. Milford. Vision-based place recognition: how low can you go? *Int. J. Rob. Res.*, 32(7), 2013.

[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.

[Nelson *et al.*, 2015] Peter Nelson, Winston Churchill, Ingmar Posner, and Paul Newman. From dusk till dawn: Localisation at night using artificial light sources. *Proc. - IEEE Int. Conf. Robot. Autom.*, 2015-June:5245–5252, 2015.

[Paul *et al.*, 2010] G Paul, S Webb, D K Liu, and G Dissanayake. A Robotic System for Steel Bridge Maintenance: Field Testing. *Proc. Australas. Conf. Robot. Autom.*, pages 1–8, 2010.

[Pepperell *et al.*, 2014] Edward Pepperell, Peter I. Corke, and Michael J. Milford. All-environment visual place recognition with SMART. In *Proc. - IEEE Int. Conf. Robot. Autom.*, pages 1612–1618, 2014.

[Rohrig and Muller, 2009] C. Rohrig and M. Muller. Localization of Sensor Nodes in a Wireless Sensor Network Using the nanoLOC TRX Transceiver. *VTC Spring 2009 - IEEE 69th Veh. Technol. Conf.*, pages 1–5, 2009.

[Rusu and Hayes, 2011] S Radacina Rusu and M J D Hayes. Localization in Large-Scale Underground Environments with RFID. *24th Can. Conf. Electr. Comput. Eng.*, pages 1140–1143, 2011.

[Salas-Moreno *et al.*, 2013] Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H.J. Kelly, and Andrew J. Davison. SLAM++: Simultaneous localisation and mapping at the level of objects. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 1352–1359, 2013.

[Sathyan *et al.*, 2011] Thuraiappah Sathyan, David Humphrey, and Mark Hedley. WASP: A system and algorithms for accurate radio localization using low-cost hardware. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 41(2):211–222, 2011.

[Sünderhauf *et al.*, 2015] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of ConvNet features for place recognition. In *IEEE Int. Conf. Intell. Robot. Syst.*, volume 2015-Decem, pages 4297–4304, 2015.

[Taylor *et al.*, 2016] Russell H Taylor, Arianna Menciassi, Gabor Fichtinger, Paolo Fiorini, and Paolo Dario. Medical Robotics and Computer-Integrated Surgery. *Springer Handb. Robot.*, pages 1657–1684, 2016.

[Wang *et al.*, 2017] Zerui Wang, Sing Chun Lee, Fangxun Zhong, David Navarro-alarcon, Yun-hui Liu, Anton Deguet, Peter Kazanzides, Russell H Taylor, and Life Fellow. Image-Based Trajectory Tracking Control of 4-DoF Laparoscopic Instruments Using a Rotation Distinguishing Marker. *IEEE Robot. Autom. Lett.*, 2(3):1586–1592, 2017.

[Zhou and Shi, 2009] Junyi Zhou and Jing Shi. RFID localization algorithms and applications-A review. *J. Intell. Manuf.*, 20(6):695–707, 2009.