Skyline-based localisation for aggressively manoeuvring robots using UV sensors and spherical harmonics

2016 IEEE International Conference on Robotics and Automation (ICRA)
Stockholm, Sweden, May 16-21, 2016

Skyline-based Localisation for Aggressively Manoeuvring Robots using UV sensors and Spherical Harmonics

Thomas Stone*[1], *Student Member, IEEE*, Dario Differt*[2], Michael Milford[3], *Member, IEEE*, and Barbara Webb[4]

***Abstract*— Place recognition is a key capability for navigating robots. While significant advances have been achieved on large, stable platforms such as robot cars, achieving robust performance on rapidly manoeuvring platforms in outdoor natural conditions remains a challenge, with few systems able to deal with both variable conditions and large tilt variations caused by rough terrain. Taking inspiration from biology, we propose a novel combination of sensory modality and image processing to obtain a significant improvement in the robustness of sequence-based image matching for place recognition. We use a UV-sensitive fisheye lens camera to segment sky from ground, providing illumination invariance, and encode the resulting binary images using spherical harmonics to enable rotation-invariant image matching. In combination, these methods also produce substantial pitch and roll invariance, as the spherical harmonics for the sky shape are minimally affected, providing the sky remains visible. We evaluate the performance of our method against a leading appearance-invariant technique (SeqSLAM) and a leading viewpoint-invariant technique (FAB-MAP 2.0) on three new outdoor datasets encompassing variable robot heading, tilt, and lighting conditions in both forested and urban environments. The system demonstrates improved condition- and tilt-invariance, enabling robust place recognition during aggressive zigzag manoeuvring along bumpy trails and at tilt angles of up to 60 degrees.**

I. INTRODUCTION

Place recognition is an important component of robot navigation [1]. Although impressive results have been demonstrated using view-based approaches (e.g. [2], [3]), many of these approaches only work under the assumption that training and test route are undertaken in similar lighting conditions. Other methods such as SeqSLAM [4] exhibit good condition-invariance but fail when viewpoint changes significantly. Many actual and potential robotic scenarios, from military reconnaissance to mowing the lawn [5], [6], involve small cheap robots moving at high speed over variable terrain, and hence revisiting places at different times with significantly different pose orientations. This paper addresses the challenge of robustly recognising these places despite rotation (yaw) change, tilt (roll, pitch) change, and appearance change.

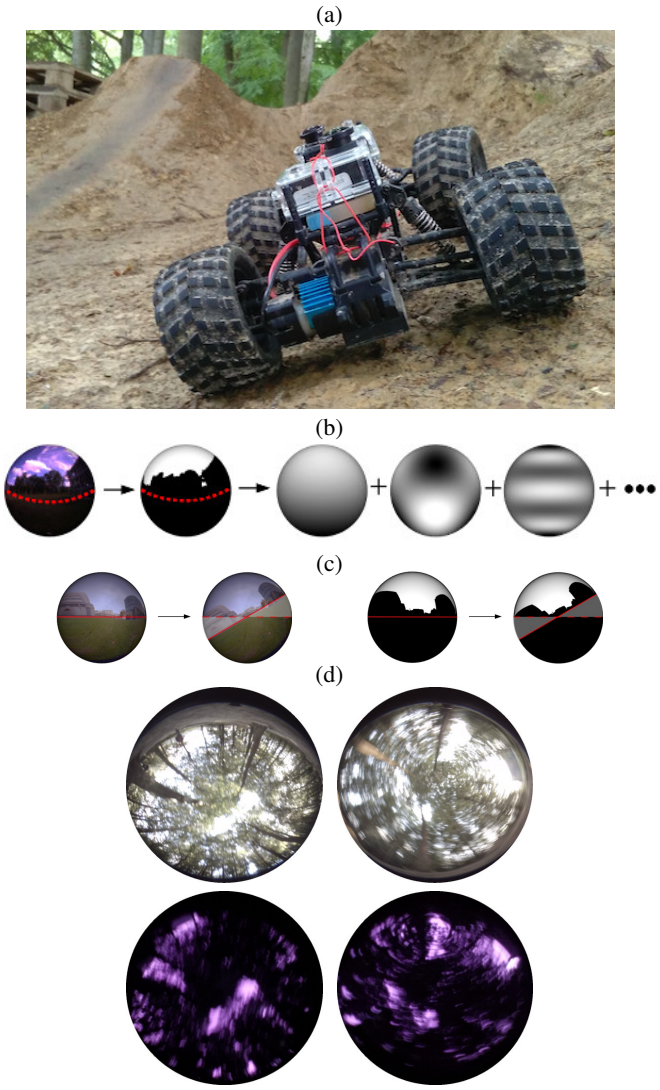

Fig. 1: (a) A high-speed all-terrain platform with UV fisheye lens oriented upwards. (b) The camera captures one hemisphere (above red line); we segment the sky, and convert it to rotation- and tilt-invariant spherical harmonic (SH) amplitudes. (c) Left: RGB images with different tilt have non-overlapping areas that need to be inferred (grey area) to match. Right: in sky-segmented images unknown areas are assumed to be ground (black) and the SH amplitudes will be unaffected by tilt. (d) Four images from the same location across two runs (one run per column, top: RGB, bottom: UV). RGB images are hard to match due to tilt, yaw rotation and motion blur; using SH amplitudes, we can correctly match the UV images.

*First author
[1]Thomas Stone is a PhD student at the School of Informatics, University of Edinburgh, 11 Crichton St, Edinburgh, Midlothian, EH8 9LE, United Kingdom t.j.stone@sms.ed.ac.uk
[2]Dario Differt is with the Computer Engineering Group, Faculty of Technology, Bielefeld University, D-33594 Bielefeld, Germany dario.differt@uni-bielefeld.de
[3]Michael Milford is with the Australian Centre for Robotic Vision at the Queensland University of Technology, Brisbane, Australia michael.milford@qut.edu.au
[4]Barbara Webb is Professor of Biorobotics at the School of Informatics, University of Edinburgh B.Webb@ed.ac.uk

Our approach is inspired by the capacity of ants to perform robust visual navigation despite substantial pitch and tilt variation in their views caused by legged locomotion over uneven terrain [7]. Ants achieve these remarkable feats of navigation by extracting skyline contours [8] exploiting UV sensitivity in the insect ommatidia [9]. We combine a novel sensing modality—a camera equipped with a UV transmitting lens—and a generalisation of the Fourier transform to the sphere known as spherical harmonics (SH) to encode the UV-segmented skyline. As described in section III, this produces descriptors that are invariant to illumination, roll, pitch and yaw, to which a sequence matching approach can be applied to enable robust place recognition along a previously traversed route. We collected novel datasets on routes with substantial condition and tilt variation (section IV), and demonstrate successful performance of our method in section V.

II. BACKGROUND

The low cost, low power consumption, and compactness of cameras has produced much interest in developing robotic navigation systems that operate from camera imagery alone. A standard approach consists of detecting features in the visual field and tracking these in subsequent views so as to reconstruct the translation and rotation of the camera relative to the incrementally expanding feature map [2], [3]. The use of rotation- and scale-invariant features makes these methods robust to different viewpoints [10]. However, they are prone to failure under different lighting conditions [11] and are also challenged in settings that lack reliable, unique features. Recent forays into using features learnt using deep convolutional networks [12] have yielded some success, but their ability to deal with appearance variation is still limited to moderate changes. These methods are also typically computationally intensive and require relatively high-resolution, high-quality imagery; constraints that are not necessarily feasible on small, low-cost robot platforms.

In contrast, methods based on global features or whole images have proven to be highly invariant to substantial lighting changes, even when using low resolution imagery [4]. However, the performance of whole image and global descriptor-based methods typically degrades rapidly when camera viewpoint changes upon multiple revisits to a place [13]. Efforts have been made to develop viewpoint invariant full-image localisation algorithms, for example, by generating multiple views for each location [14]. This solves particular problems such as the translation due to lane position, but it soon becomes intractable to generate views that deal with all types of pose invariance due to roll, tilt and yaw, especially if computational resources are limited. For less agile robot platforms, this problem can be partially addressed by the use of accelerometers that can compensate for tilt on the fly by remapping the view. However, a gravity vector cannot be reliably extracted from an accelerometer on a platform that is regularly accelerating and decelerating [15], and the use of high sampling rate inertial measurement units (IMUs) adds substantial expense.

A potential image-based solution to this problem is to use an omnidirectional camera and a compact image encoding that is pose invariant. In principle, lower order SH can be used in this way by converting an image to greyscale, projecting it onto a hemisphere and filling in the missing areas through interpolation or reflection before calculating the coefficients [16]. However, using SH precludes the ability to incorporate the methods used to ameliorate illumination variation, such as local histogram equalisation on neighbouring pixels (patch normalisation) used in vanilla SeqSLAM [4]. Moreover, unless the full spherical view can be captured, which is difficult for any practical robot sensor configuration, this method can also fail to handle robot pitch and roll. For example, if only the upper hemisphere is captured, then even small camera pose variations can cause substantial differences between two images and their harmonics (fig. 1c). Detecting and correcting for tilt may not suffice because different areas of the view will be missing and would need to be interpolated.

Both limitations can potentially be overcome by preprocessing the images to extract an illumination invariant feature that remains visible in a hemispherical view under tilt and yaw. Such a feature in outdoor scenes is the boundary between sky and ground. The shape of the sky at a particular location can provide a unique signature [17], [18]. Due to low reflectivity of most objects in the ultraviolet (UV) range [19], UV sensors are particularly suitable for segmenting the sky from images with minimal computational cost [20], [21]. We have previously shown that sky extraction using a UV-adapted camera can be successfully combined with SeqSLAM for navigation, albeit on a large and stable platform traversing flat city streets [22]. Here we develop a UV-based tilt-invariant place recognition system by combining this novel sensing modality with the use of SH.

III. APPROACH

In this section we describe the novel method for extracting skylines using a UV-sensitive camera, and encoding a tilt-, rotation- and illumination-invariant snapshot of each place the robot visits based on SH, which can then be fed into a sequence-based place recognition algorithm.

A. Skyline Extraction using UV

Many natural materials show only a low reflectivity to light with short wavelengths, while the light emitted by the sky contains a comparably high portion of UV light [21]. Sensors which are only sensitive to shorter wavelengths will therefore perceive a high contrast between ground objects and the sky [22]. We use a camera with a UV-sensitive lens (see section IV) to extract a skyline by taking an average of all channels in the image and applying a threshold using Otsu's method [23]. The resulting black/white image is invariant to changes of the lighting conditions as long as a classification between the sky and ground is possible. Moreover, when two binary hemispherical images from the same location are tilted differently, we can safely assume that the non-overlapping parts of the image are largely black, due to the

natural position of the sky falling well within the upper hemisphere (fig. 1c). Once a binary sky shape has been extracted, matching it to a previously seen view reduces to a 2D shape recognition problem. Solutions to translation- and rotation-invariant shape recognition problems exist (e.g. [24], [25]). However, as our sky is mapped on a 3D sphere we need to apply a suitable basis that can be used for recognition of shapes on spheres, such as SH [26].

B. Real Spherical Harmonics

The generalisation of the Fourier transform to the sphere is a well-studied topic in the field of Fourier analysis [27], [28]. While the standard approach is to use complex-valued SH, instead we use the real-valued equivalent, real spherical harmonics (RSH), to avoid computations with complex numbers. Since both bases differ only in their representation we will in the following denote them by SH. Fourier analysis states that there exists a base of real-valued orthogonal functions $y_m^l : S^2 \to \mathbb{R}$ on the sphere with

$$y_m^l(\vartheta,\varphi) := \begin{cases} \sqrt{2}N_m^l \cos(m\varphi)P_m^l(\cos\vartheta) & m > 0 \\ \sqrt{2}N_m^l \sin(-m\varphi)P_{-m}^l(\cos\vartheta) & m < 0 \\ N_0^l P_0^l(\cos\vartheta) & m = 0 \end{cases} \quad (1)$$

and $l \in \mathbb{N}, -l \leq m \leq l$, where l denotes the *band* of the corresponding SH, P_m^l are the *associated Legendre polynomials* and $N_m^l = \sqrt{\frac{2l+1}{4\pi}}\sqrt{\frac{(l-|m|)!}{(l+|m|)!}}$ is a normalisation term. Note that we used spherical coordinates (ϑ, φ) in (1). For an efficient implementation the values $y_m^l(\vartheta,\varphi)$ can be calculated by using recursive formulas, e.g. [29]. Any real-valued function $f : S^2 \to \mathbb{R}$ can be expressed in terms of SH by a projection

$$c_m^l := \langle f, y_m^l \rangle = \int_S f(s) y_m^l(s) ds \in \mathbb{R}, \quad (2)$$

such that $f \approx \sum_{l,m} c_m^l y_m^l$ converges quadratically in $L^2(S^2)$ with an increasing value of l. Therefore each function f can be approximated by l^2 coefficients c_m^l by using the first l bands of the SH.

By defining the amplitude spectrum on the SH as

$$a^l := \sqrt{\sum_{m=-l}^{l} (c_m^l)^2}, \quad (3)$$

a rotational-invariant measure can be deduced analogously to the amplitude spectrum in the standard Fourier transform. It contains a total of l entries for a function approximated up to band l. Due to the missing phase information, the total information contained in the amplitude spectrum is less then the information provided by the function f.

C. Place Recognition Algorithms

SeqSLAM is a place recognition algorithm that uses whole images as global features [4] and has been shown to be effective using very low-resolution, blurred images [30]. It works by repeatedly searching previously stored images for locally good matches with the current image. Linear sequences of candidate images are interpreted as a high probability of being in the corresponding location. SeqSLAM can be applied to feature vectors as well as images, and we here apply it to vectors of SH amplitude coefficients, and compare this to vanilla SeqSLAM (i.e., using greyscale images). In the following we will denote the original by *SeqSLAM (vanilla)* and our approach by *SeqSLAM (sky)*.

OpenFABMAP is an implementation of FAB-MAP 2.0 that uses feature detection to build a visual vocabulary [2]. This is then used for a bag-of-words [31] approach to approximate the likelihood of a particular frame matching the current view. Since FAB-MAP is a feature-based method it is invariant to changes in the viewpoint but encounters problems if the feature detection is disturbed (blur, lighting changes, or featureless environments).

IV. EXPERIMENTAL SETUP

In order to evaluate the robustness of the proposed technique to deal with tilt and appearance variation in real environments, we first conducted a diagnostic tilt-controlled experiment, then evaluated the system in two challenging robot scenarios. See attached video for details.

A. Cameras and Rigs

To capture our datasets we used two identical cameras (GoPro Hero 3+ Black) mounted with different lenses. The first camera had a RageCams $1/3.2''$ panoramic fisheye lens with a focal length of $1.19\,\mathrm{mm}$, $185°$ field of view and an infrared (IR) cut filter, which is able to collect RGB images. The second camera is mounted with a similar, but modified, fisheye lens provided by skylinesensors.com. The lens has an additional custom IR cut filter and UV bandpass filter, with a peak response around $350\,\mathrm{nm}$. Raw camera data was recorded with the same settings on both (Video resolution "1080p super view", 24 FPS, wide field of view, automatic exposure time). The only significant cost for making this camera sensitive to the UV spectrum was the filter. Currently, a one-off filter with custom response costs several hundred dollars, but if mass produced would cost far less. Costs could be further reduced by mass producing low pixel count, small sensor arrays sensitive to UV.

The cameras were used in two different experimental rigs (fig. 2). The first rig consisted of each camera mounted (one at a time) on the top of a helmet with an adjustable tilting angle α (fig. 2a) to record a ride on a bicycle. The second rig consisted of an off-the-shelf remote control car (RCC) (Drive&Fly-models TruckFighter) with both cameras mounted next to each other on the top of the RCC, replacing the plastic chassis (fig. 2b).

B. Datasets

The City dataset (fig. 3a, 4a) formed the basis for controlled tilt evaluation. It was collected using the bike helmet rig on a circular track of $520\,\mathrm{m}$ along urban streets in Bielefeld at around 3pm. The route was traversed eight times with varying tilting angles $\alpha \in (-30, -20, \ldots, 30)$—where $\alpha = 0$ has been recorded twice—and with each camera

(a) Variable tilt

(b) RCC

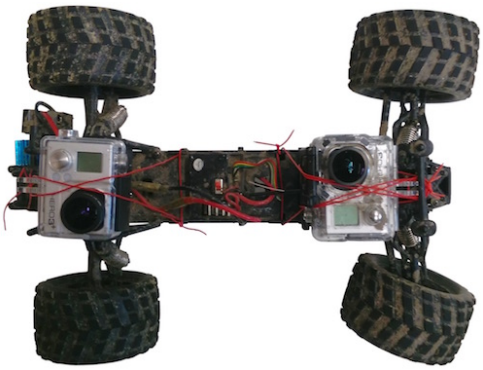

Fig. 2: (a) Side-view of the bicycle helmet setup with a single camera mounted at each run. The additional tilting is denoted by the angle α, where $\alpha = 0$ with the camera approximately level. (b) Top-view of the RCC platform mounted with the RGB and UV cameras simultaneously, both pointing upwards.

(a) City (variable tilt) (b) BMX Track (RCC) (c) Disposal Site (RCC)

Fig. 3: Aerial imagery of the three different testing environments, with the red arrow line indicating the route, traversed at least twice for each experiment. Map data: Google, Geobasis-DE/BKG.

(a) City (variable tilt) (b) BMX Track (RCC) (c) Disposal Site (RCC)

Fig. 4: Photos of the three testing environments taken with a regular digital camera, showing the wide range of urban, forested and industrial areas encountered.

(RGB, UV), totalling 16 runs. Over the course of all the runs, substantial scene dynamics were encountered including those due to varying lighting conditions and traffic.

The other two data sets were gathered using the RCC rig, on two different days, at different times of day, to increase the visual dynamics (e.g. obstacles, illumination) of the observed scenes. The BMX Track data (fig. 3b, 4b) was collected in an area close to Bielefeld university and contains a high diversity of visual cues. The $600\,\mathrm{m}$ route starts on a bicycle motocross (BMX) track in a forest, containing sloping terrain and a heavy forest canopy with minimal unique landmarks. The route then leaves the forest and continues along a border between a forest (right side) and the gardens of adjoining houses (left side). Finally, the route ends along a suburban street. The Disposal Site dataset (fig. 3c, 4c) was recorded on a highly repetitive track in Borgholzhausen, presenting potential aliasing issues. The $510\,\mathrm{m}$ track consists mainly of a road through a forest marked by a single road junction, a disposal site and a small building.

To ground truth the datasets, key corresponding frames in all environment traverses were visually identified. Image pairs between these key frames were matched through linear interpolation. When calculating precision, images were deemed a correct match when the estimated frame correspondence fell within 5 frames from the ground truth.

C. Parameter Values

OpenSeqSLAM was run using all default parameters: $ds = 10$, $v_{min} = 0.8$, $v_{max} = 1.2$, $r_{window} = 10$. For OpenFABMAP, we trialled a range of different detectors used at varying resolutions. For best performance we used 480×270 pixel images with Speeded Up Robust Features (SURF) [10] detectors and extractors to build a visual vocabulary, using training data supplied from a separate video dataset recorded in similar environments for both the BMX Track and Disposal Site datasets. All other parameters were kept as provided in the latest sample script from the OpenFABMAP public repository (accessed Aug. 2015).

For SeqSLAM (sky) we used 2000 pixels to sample the upper hemisphere from the UV images and Otsu's method for the binarisation which does not depend on any parameters. A total of $l = 120$ bands were used for the creation of the feature vector containing a total of 14400 entries per input image (equivalent to the data contained in a square 120×120 pixel monochromatic image).

V. RESULTS

A. City Tilt Variation

As can be seen in figure 5b, place recognition performance using SeqSLAM (vanilla) and SeqSLAM (sky) is broadly comparable at low tilt angles, but diverges rapidly as the tilt angle increases. At a tilt difference of $60°$, a recall of nearly $20\,\%$ can still be achieved with a precision of $100\,\%$ using the UV images. The vanilla method is unable to correctly match any images at a similar precision level, even at $50°$. Figure 6 shows an example of two frames that were successfully matched using the sky segmented imagery.

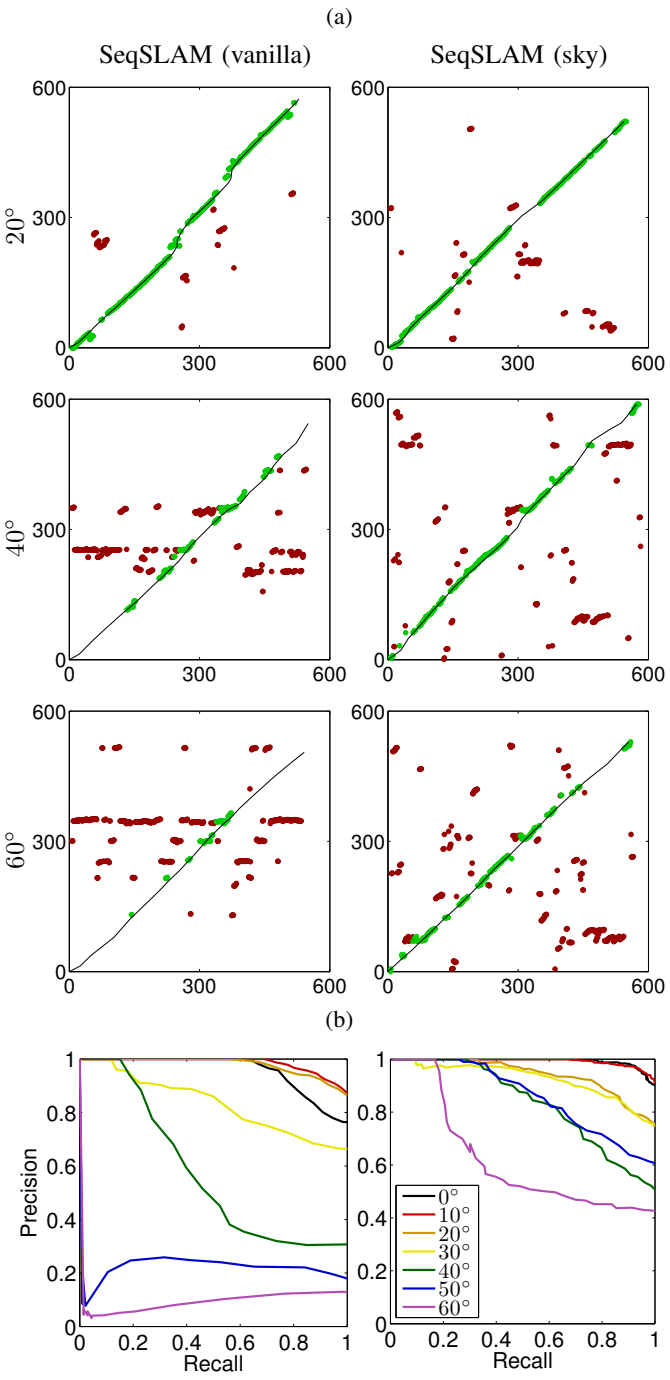

Fig. 5: (a) Frame correspondences (place matches between the test and training environment traverses) on the City tilt set at $20°$, $40°$, and $60°$ tilt differentials. Left: vanilla SeqSLAM. Right: improved performance achieved with sky SeqSLAM, both in terms of total number of correct matches and matching coverage throughout the dataset. Solid line: ground truth; Green dots: near ground truth; Red dots: far from ground truth; Axis units: frame indices. (b) Precision versus recall at various degrees of tilt. Runs were captured at values of α between $-30°$ and $30°$ and compared by pairing $\alpha_1 \approx -\alpha_2$. As the difference in tilt between runs increases, performance degrades more slowly for sky SeqSLAM than for vanilla SeqSLAM.

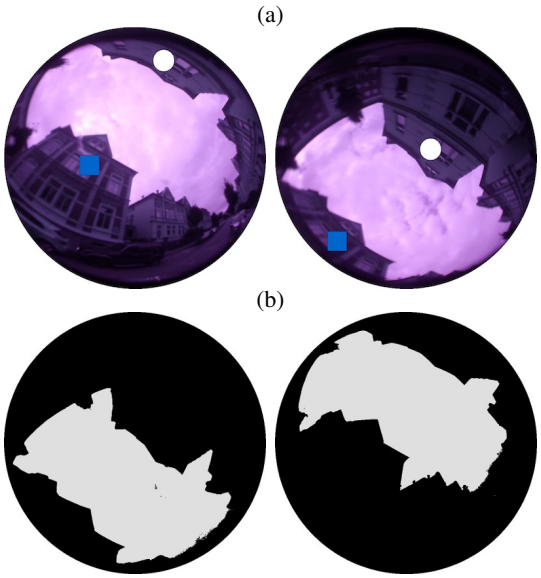

Fig. 6: (a) Two hemispherical UV images taken at the same location with a 60° tilt offset from each other. The white circle and blue square show corresponding parts of the images. Distortion and misalignment results in poor place recognition performance with vanilla SeqSLAM. (b) Extracting the shape of the sky gives an excellent match using SH amplitudes.

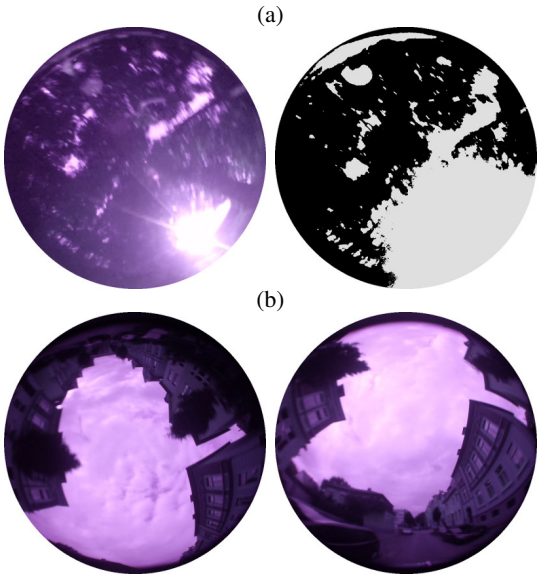

Fig. 7: Failure cases: (a) (left) An image with bright direct sunlight in a dark area causing high amounts of lens flare, leading to (right) incorrect segmentation of sky and ground. (b) Strongly tilted views recorded at the same position, which can not be matched due to cropping of the UV segmented skyline image.

The difference in matching performance can also be examined in the frame correspondence plots shown in figure 5a. At tilt angles of 20°, 40° and 60°, the UV-based method finds more true positive place matches than the vanilla SeqSLAM method. Place recognition coverage throughout the route is also improved.

B. BMX Track

As described in section IV and as shown on the map in figure 3b, the BMX Track dataset consists of three sections which strongly differ in their visual appearance. This particularly affects the correspondences found by FAB-MAP (fig. 8a), which can mostly distinguish between the three different areas on the track but struggles to find the RCC's exact location, with precision no better than 50 % for most recall levels (fig. 8d). In comparison, both SeqSLAM methods show good performance on the complete BMX track at 100 % precision, with recall rates of 33 % and 37 % for the vanilla and UV segmented skyline approach, respectively. Vanilla SeqSLAM tends to fail on parts of the track that include misaligned images due to either tilt or rotation. The precision levels achieved while maintaining 100 % recall (the complete dataset) are 68 % for vanilla and 83 % for sky SeqSLAM respectively, showing superior localisation with the UV segmented skyline. One problem we encountered on this dataset is lens flare, leading to errors during the skyline segmentation as can be seen in figure 7a. However, we found that the restricted view of the sky through foliage (fig. 1d) did not impact the performance of SeqSLAM using UV segmented skyline images, although no clear skyline can be seen. Instead, the spots of sky visible through the canopy were sufficient on our dataset to localise the RCC.

C. Disposal Site

On the Disposal Site all algorithms performed worse, probably due to the highly repetitive nature of the track (fig. 9d). FAB-MAP failed catastrophically using the best performing parameters we could find (fig. 9a). Vanilla SeqSLAM failed due to a combination of zigzagging (rotational differences) and the strongly repetitive visuals, achieving poor frame correspondence towards the end of the route (fig. 9b) and a precision level of 33 % at 100 % recall. Using the sky does not completely fix this problem as there is no viewpoint invariance for translations across the x-y plane. Despite this our algorithm performed well (fig. 9c) with 64 % precision at 100 % recall.

D. Tilt Invariance versus Sequence Length

As indicated in figures 8d and 9d the precision rates can be improved by increasing the sequence length used in sky SeqSLAM. With a sequence length of 20, we obtain a precision level of 89 % for the BMX track and 70 % for the Disposal Site at 100 % recall. A comparison using single images (rather than sequences) shows a surprisingly high precision level of 64 % and 49 % at 100 % recall over the whole dataset, despite the fact that the amplitude spectrum generally provides ambiguous information for images.

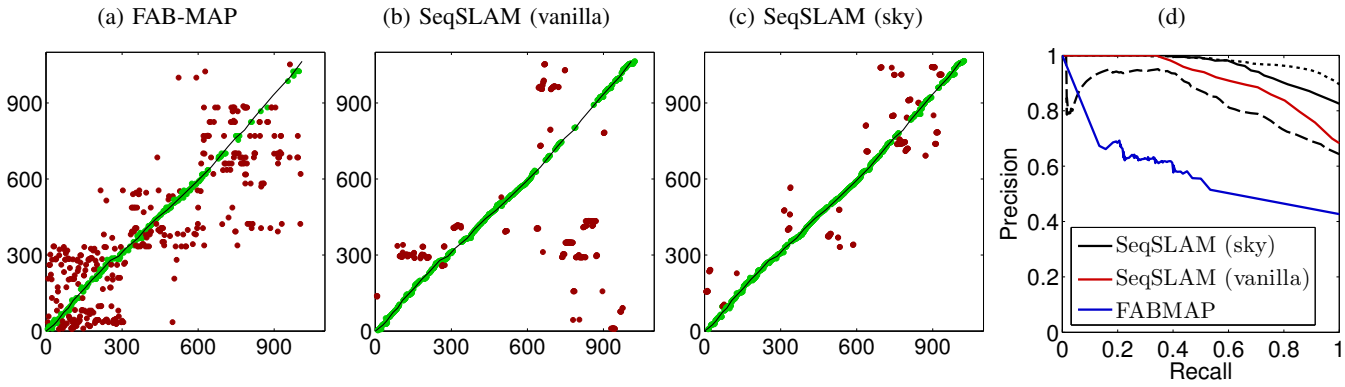

Fig. 8: (a-c) Frame correspondences on the BMX Track dataset. Using UV sky segmentation improves the matching performance over both FAB-MAP and vanilla SeqSLAM. (d) Precision vs. recall. SeqSLAM (sky) performs best, and can be improved by increasing the sequence length from the default of $w = 10$, which equates to 2 seconds of travel, to $w = 20$ (dotted line). Dashed line is ($w = 1$).

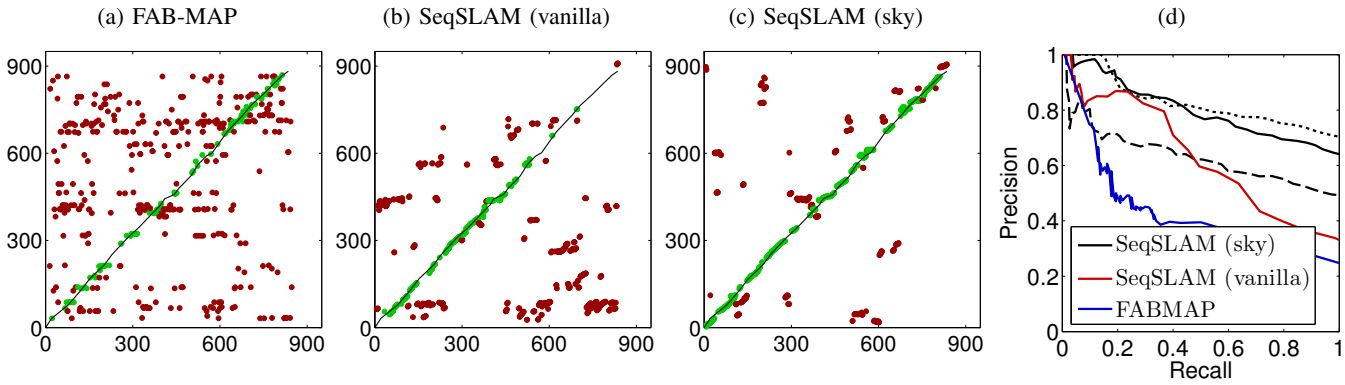

Fig. 9: (a-c) Frame correspondences on the Disposal Site dataset. As well as improving overall matching performance, the UV sky segmented implementation particularly improved performance at the end of the dataset where aggressive zig-zag manoeuvring was performed. (d) Precision vs. recall. Performance of adjusted sequence length using UV is shown by dashed ($w = 20$) and dotted ($w = 1$) lines, as in fig. 8d.

E. Computational Efficiency

In our implementation, the computation of the Fourier transform is $O(mn^2)$ and the amplitude spectrum $O(n^2)$, where m is the number of sample points and n the number of bands. For the presented results we used $n = 120$ bands and $m = 2000$ visual points equally distributed over the sphere. The run-time for both the Fourier transform and the calculation of the amplitude spectrum is around $40\,\mathrm{ms}$ on a Lenovo T-530 laptop. An additional speed-up could be gained by reducing the number of used bands and/or visual points.

VI. DISCUSSION AND FUTURE WORK

In this paper we have presented a novel method for place recognition that combines UV-sensitive camera imaging with a compact tilt- and rotation-invariant skyline representation using SH amplitudes. The proposed system is particularly relevant for low-cost, high-speed robotic platforms operating over challenging terrain and across changing environmental conditions. The UV-modification can be cheap and provides appearance robustness and the SH representation provides invariance to the major viewpoint change challenges faced by these platforms. This can be exploited for use in a sequence-based matching algorithm, which would not work on tilt-sensitive imagery. The controlled tilt and robot platform experiments presented here demonstrate that the approach is feasible over a wide range of environments and platforms, and performs significantly better than current state of the art condition-invariant (SeqSLAM) and viewpoint-invariant (FAB-MAP) approaches. While SH are a natural choice for the stated problem, other shape recognition techniques could be adapted to work on a sphere. Here we discuss some of the shortcomings of this approach and promising areas of future work.

In large flat areas such as a desert, the use of a skyline feature is less beneficial, both due to homogeneity of the skyline and the introduction of skyline distortion at a smaller tilt angle than in wooded or urban environments where the skyline is high. These shortcomings could be somewhat mitigated if a greater than $180°$ sensor was used, such as the Ricoh M15, a full $360° \times 180°$ dual fisheye camera.

Like all camera-based methods, our method is sensitive to

extreme lens flare, such as that caused by direct sunlight very early or late in the day. In such situations, ground pixels can be labelled as sky during segmentation, leading to degraded place recognition performance. This problem could potentially be overcome in a number of ways; firstly by semantic recognition of similar situations and a consequent downgrading of the place recognition confidence, or secondly by attempting to match snapshots using a mask that excludes or adapts areas of maximum brightness in the image, similar to the approach in [20].

Although we have applied this system on high-speed ground-based platforms, it could also have utility on aerial platforms which can rapidly change their attitude when moving. The method would have particular advantages in complex environments such as cities and forests.

The use of UV and skyline for navigation was inspired by research on the behaviour and visual system of ants. While it is not explicitly known how ants represent or process this information, a basis encoding such as SH is a possibility that would be consistent with current hypotheses for visual processing in other animals [32]. Investigating the underlying neural mechanisms in the insect brain could lead to further insight into compact and low cost computation for navigation.

ACKNOWLEDGMENT

This work was supported by a Scottish Informatics and Computer Science Alliance Distinguished Visting Fellow grant, Research Council Future Fellowship FT140101229 and Microsoft Research Faculty Fellowship to Michael Milford. We thank the Deutsche Forschungsgemeinschaft (DFG) for financial support (grant numbers MO 1037/8-1 and MO 1037/10-1), and also acknowledge grants EP/F500385/1 and BB/F529254/1 for the University of Edinburgh School of Informatics Doctoral Training Centre in Neuroinformatics and Computational Neuroscience (www.anc.ac.uk/dtc).

REFERENCES

[1] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 2. IEEE, 2000, pp. 1023–1029.

[2] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0." in *Robotics: Science and Systems*, vol. 5. Seattle, USA, 2009.

[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1052–1067, 2007.

[4] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2012, pp. 1643–1649.

[5] L. Matthies, Y. Xiong, R. Hogg, D. Zhu, A. Rankin, B. Kennedy, M. Hebert, R. Maclachlan, C. Won, T. Frost, *et al.*, "A portable, autonomous, urban reconnaissance robot," *Robotics and Autonomous Systems*, vol. 40, no. 2, pp. 163–172, 2002.

[6] J. Yang, S.-J. Chung, S. Hutchinson, D. Johnson, and M. Kise, "Omnidirectional-vision-based estimation for containment detection of a robotic mower," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2015.

[7] P. Ardin, M. Mangan, A. Wystrach, and B. Webb, "How variation in head pitch could affect image matching algorithms for ant navigation," *Journal of Comparative Physiology A*, vol. 201, no. 6, pp. 585–597, 2015.

[8] P. Graham and K. Cheng, "Ants use the panoramic skyline as a visual cue during navigation," *Current Biology*, vol. 19, no. 20, pp. R935–R937, 2009.

[9] R. Möller, "Insects could exploit UV–green contrast for landmark navigation," *Journal of Theoretical Biology*, vol. 214, no. 4, pp. 619–631, 2002.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[11] C. Valgren and A. J. Lilienthal, "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments," *Robotics and Autonomous Systems*, vol. 58, no. 2, pp. 149–156, 2010.

[12] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," *Proceedings of Robotics: Science and Systems XII*, 2015.

[13] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. of Workshop on Long-Term Autonomy, International Conference on Robotics and Automation (ICRA)*. IEEE, 2013, p. 2013.

[14] M. Milford, C. Shen, S. Lowry, N. Sünderhauf, S. Shirazi, G. Lin, F. Liu, E. Pepperell, C. Lerma, B. Upcroft, and I. Reid, "Sequence searching with deep-learnt depth for condition- and viewpoint-invariant route-based place recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2015.

[15] S. Thurrowgood, D. Soccol, R. J. Moore, D. Bland, and M. V. Srinivasan, "A vision based system for attitude estimation of UAVs," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2009, pp. 5725–5730.

[16] H. Friedrich, D. Dederscheck, M. Mutz, and R. Mester, "View-based robot localization using illumination-invarant spherical harmonics descriptors," in *Proceedings of the International Joint Conference on Computer Vision and Computer Graphics Theory and Applications (VISAP)*. INSTICC, 2008.

[17] F. Stein and G. Medioni, "Map-based localization using the panoramic horizon," in *International Conference on Robotics and Automation*. IEEE, 1992, pp. 2631–2637.

[18] J.-i. Meguro, T. Murata, Y. Amano, T. Hasizume, and J.-i. Takiguchi, "Development of a positioning technique for an urban area using omnidirectional infrared camera and aerial survey data," *Advanced Robotics*, vol. 22, no. 6-7, pp. 731–747, 2008.

[19] T. Kollmeier, F. Röben, W. Schenck, and R. Möller, "Spectral contrasts for landmark navigation," *Journal of the Optical Society of America A*, vol. 24, no. 1, pp. 1–10, 2007.

[20] M. H. Tehrani, M. Garratt, and S. Anavatti, "Horizon-based attitude estimation from a panoramic vision sensor," in *Embedded Guidance, Navigation and Control in Aerospace*, vol. 1, no. 1, 2012, pp. 185–188.

[21] D. Differt and R. Möller, "Insect models of illumination-invariant skyline extraction from UV and green channels," *Journal of theoretical biology*, vol. 380, pp. 444–462, 2015.

[22] T. Stone, M. Mangan, P. Ardin, and B. Webb, "Sky segmentation with ultraviolet images can be used for navigation," in *Proceedings Robotics: Science and Systems*, 2014.

[23] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[24] M.-K. Hu, "Visual pattern recognition by moment invariants," *Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[25] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 489–497, 1990.

[26] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," in *Symposium on geometry processing*, vol. 6, 2003, pp. 156–164.

[27] W. E. Byerly, *An elementary treatise on Fourier's series and spherical, cylindrical, and ellipsoidal harmonics, with applications to problems in mathematical physics*, 1st ed. London, England: Ginn & Company, 1893.

[28] G. B. Folland, *Fourier analysis and its applications*, 1st ed. Providence, Rhode Island: American Mathematical Society, 1992.

[29] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical recipes in C: The art of scientific computing*, 2nd ed. Cambridge, England: Cambridge University Press, 1992.

[30] M. Milford, "Vision-based place recognition: How low can you go?" *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 766–789, 2013.

[31] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 977–984.

[32] B. A. Olshausen *et al.*, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.