Shady Dealings: Robust, Long-Term Visual Localisation using Illumination Invariance

Colin McManus, Winston Churchill, Will Maddern, Alexander D. Stewart, and Paul Newman

***Abstract*— This paper is about extending the reach and endurance of outdoor localisation using stereo vision. At the heart of the localisation is the fundamental task of discovering feature correspondences between recorded and live images. One aspect of this problem involves deciding *where* to look for correspondences in an image and the second is deciding *what* to look for. This latter point, which is the main focus of our paper, requires understanding how and why the appearance of visual features can change over time. In particular, such knowledge allows us to better deal with abrupt and challenging changes in lighting. We show how by instantiating a parallel image processing stream which operates on illumination-invariant images, we can substantially improve the performance of an outdoor visual navigation system. We will demonstrate, explain and analyse the effect of the RGB to illumination-invariant transformation and suggest that for little cost it becomes a viable tool for those concerned with having robots operate for long periods outdoors.**

I. Introduction

Feature-based stereo vision localisation can be simply understood as the act of matching run-time observed visual features to stored features and then estimating the pose of the vehicle given these associations. As ever, the detail is devilish. While the matching problem is indeed simply stated, its execution can be fraught with difficulty. Two problems dominate: where should one search for correspondences (and how big should a search window be) and what should one search for (what does the feature look like). We take both these issue in turn. In Section II-B, as a precursor to what follows, we offer a simple feed-forward approach to the first "spatial" component of the matching task. This results in improved data association results by way of the directed feature matching search that ensues. The substantial contribution of this paper comes in Section III onwards in where we deal with the "appearance" component of the matching task. Here we consider the effect of change in illuminant and construct an imaging pipeline, which, by virtue of modelling the effect of black body radiation on outdoor scenes, allows us to demonstrate superior localisation performance.

A. Appearance Change From Illumination

For vision systems concerned with localising in known environments, dealing with appearance changes, either sudden or gradual, is an ongoing challenge. Appearance changes can result from several sources, such as (i) different lighting conditions, (ii) varying weather conditions, and/or (iii) dynamic objects (e.g., pedestrians or vehicles). In previous work [1], we demonstrated how to leverage knowledge of prior 3D structure to suppress distracting objects for improved pose estimation in busy urban environments, as well as how to cope with long-term appearance variation caused by changing weather conditions [2]. In this paper, we attempt to address problem (i), and examine how to localise despite stark changes in lighting.

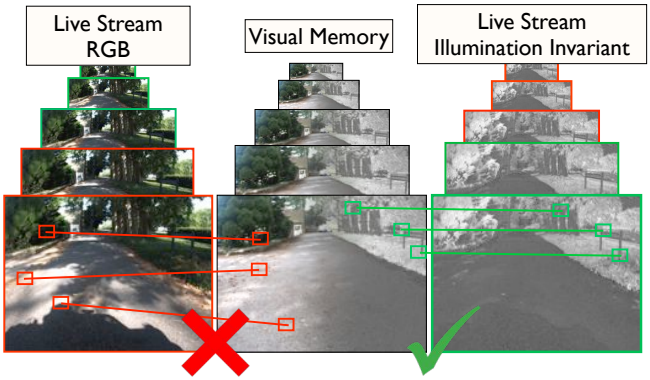


Fig. 1. We present an approach to localisation that runs two localisers in parallel — one using images in RGB colour space and the other using images in an illumination-invariant colour space. This allows us to cope with areas that exhibit a significant amount of lighting variation. The top left image sequence represents the live video stream in RGB colour space, the right image sequence represents the live video stream in an illumination-invariant colour space, and the middle sequence represents the images in our visual memory, which include both the RGB and illumination-invariant space. In our technique, if one of the localisers fails, we switch to the other.

The primary challenge presented by lighting changes are the shadows they cast, which can obscure features in the environment and create new ones from the silhouettes, making it difficult to match features from a sunny day to a cloudy day (see Figure 1). To address this problem, we leverage recent work in the computer vision field for transforming RGB-coloured images into an illumination-invariant colour space [3]. The ability to determine the colour of objects irrespective of an external illumination source is known as colour constancy [4].

We present an approach that runs two localisation threads in parallel, one using the original RGB images and the other using the illumination-invariant images. The system switches between the two estimates depending on the quality of the respective estimates. We demonstrate on over 10 km of data that this incredibly simple addition to the standard vision pipeline can result in significant improvements in areas that exhibit a great deal of lighting variation.

Authors are members of the Mobile Robotics Group, University of Oxford, Oxford, England; {colin,winston,wm,alexs,pnewman}@robots.ox.ac.uk

B. The Literature

Shadows, insufficient lighting, and changing lighting conditions has been studied in a variety of fields with different goals in mind. In this work we draw on the optics community who have paid careful attention in modelling the image formation process, considering properties of the illuminant, the camera, and the scene. Of particular relevance to this work, the optics literature shows how full colour images can be mapped to an illumination-invariant space. Finlayson's *et al.* [5][6] mapping can be computed by analysing an image in which a material property is viewed under different lighting conditions (e.g., the ground in sun and shade). Ratnasingam *et al.* [7][3] instead use known properties of the camera to produce the invariant image.

In the computer vision community, the detection and removal of shadows has been performed using learnt classifiers. Guo *et al.* [8] use a graph-cut framework involving image patches to remove shadows from natural scenes. Zhu *et al.* [9] are able to classify shadows in greyscale images using boosting and conditional random fields. Kwatra *et al.* [10] use an information theoretic method—a hybrid of the classifier and physics based approaches—to remove shadows in aerial imagery. While the results are effective, the process is relatively slow for typical image sizes.

Within the robotics community, the issues of lighting in different problems have been tackled in a variety of ways. The SeqSLAM [11] algorithm is able to achieve successful topological localisation despite extreme variations in lighting. The approach exploits the fact that sufficiently long sequences of images are distinctive enough for localisation, and they are able to localise at night against a daytime map. Corke *et al.* [12] apply Finlayson's invariant image to the problem of single-image localisation to deal with the issue of shadows. They show that the transformed images of a location were more similar than the original colour images and therefore localisation performance improved. Maddern *et al.* [13] show that place recognition can be improved over a day-night cycle by using both a standard and thermal camera; however this required specialist hardware. McManus *el al.* [14] improve the robustness of their visual teach and repeat system to lighting issues by using a lidar-based sensor, which is lighting invariant. While the sensor produces good results, it has a range of issues including cost, fragility, power requirements, frame-rate (2 Hz) and availability. The experience based navigation work by Churchill *et al.* [2] attempts to solve the lighting problem by capturing the different visual modes of an environment with different experiences. However this was found to break down when lighting effects cause new visual patterns on every visit to a location, such as shadows cast by foliage.

In this work we look to leverage the invariant image transform proposed by Ratnasingam *et al.* [3] to improve metric localisation performance and robustness in the face of strong and changing shadows, which caused the experience based navigation system to fail [2].

II. Preliminaries

A robust localisation system should be able to answer the following questions at all times: where should I look (spatially in the image) and what should I look for (appearance in the image)? We address both of these questions in coming sections and then present our approach to combining lighting-invariant images with our baseline system (RGB only). But, before we do that, we are well served by a synopsis of the underlying mechanisms we use for localisation.

A. Stereo Localisation

At the heart of our localisation system is a keyframe-based visual odometry (VO) pipeline. At runtime, we use the FAST [15] detector on both stereo images for feature extraction. We then find stereo correspondences using patch-based matching, and compute BRIEF descriptors [16] for each stereo measurement. We also compute a 3D estimate of the feature position relative to the camera frame. When a new stereo frame is acquired, features are extracted and matched to the previous frame, initially with BRIEF matching, and then refined using patch-based matching to achieve sub-pixel correspondences. RANSAC is used for outlier rejection, followed by the nonlinear solve to produce the frame-to-frame transformation estimate.

We use a survey vehicle to collect multiple stereo sequences of an environment and the output from the VO process—a series of keyframes with features locations, descriptors, pixel patches, 3D landmarks, and a relative transformation estimates—are saved in memory (these are the survey keyframes). Then, to localise the current live image sequence to a memory, we use the a similar VO pipeline as the one described above. Instead of matching to the previous camera frame, we match to a survey keyframe.

B. Knowing Where To Look

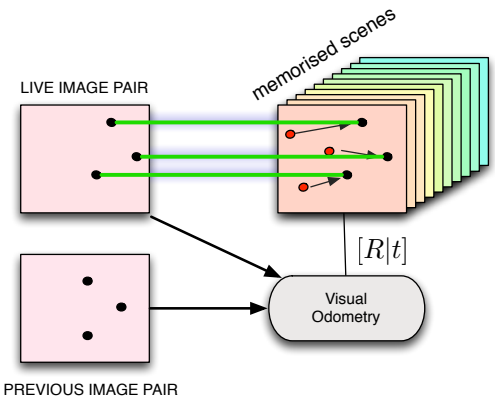


Fig. 2. Illustration of our feature prediction approach. Using the latest VO output from the live image stream, we can predict where features in the live frame should reproject in the survey keyframe. This allows us to restrict the search space for candidate feature matches, which has two benefits: (i) improves efficiency, and (ii) reduces spurious matches that result from global matching in descriptor space.

In an effort to improve robustness in matching a live view with a survey view, we take an active searching approach

similar to Davison *et al.* [17], which predicts how the measurements in the survey frame should reproject in the live frame. However, in Davison's work, they were limited to predicting the motion using a constant-velocity motion assumption. In our localisation system, we have access to the VO output from the live image stream. This allows us to accurately predict how we have moved relative to our survey and therefore inform where we expect to find stored features in the live view. Specifically, using the uncertainty in the map, measurements, prior pose estimate, and latest VO estimate, we can compute the covariance of the reprojected measurements from the survey frame into the live frame. This in turn can be used to define a search region in the live view. This is illustrated in Figure 2.

By using this active search approach [17], we are able to better predict our search regions and thus, reduce the likelihood of bad data associations. The next step, which is the key contribution to our approach, is to identify *what* to look for within each of these regions. Standard methods would attempt patch-based matching or descriptor-based matching on the raw images. However, this approach is obviously inadequate under extreme lighting changes. In the next section, we will show how a simple image transformation can help improve localisation in areas with significant lighting variation.

III. Knowing What To Look For Whatever the Lighting

Fig. 3. Example images taken around our Begbroke Science Park test site, with the raw RGB image shown on top, and the corresponding lighting invariant version shown below. Note how the image transformation is able to significantly reduce the impact of the shadows.

Given a search region for a potential match, our baseline system finds the sub-pixel location that minimises the score between the reference patch from the survey and the live image. However, as illustrated in Figure 1, this approach can fail when the appearance change is too significant. To remedy this problem, we wish to inform our system about the illuminate-free appearance of the scene, which requires a transformation from the standard RGB colour space.

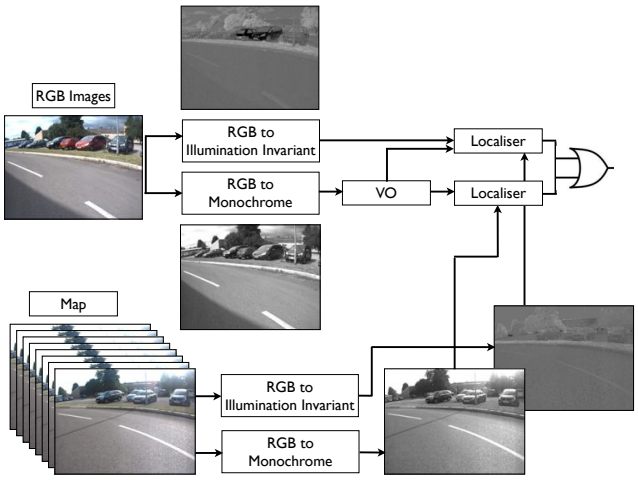


Fig. 4. Block-flow diagram of our combined localisation approach. Note how we can feed the VO estimates from the untransformed images to help predict the feature locations in the transformed images. Note that our baseline system does not work on the raw RGB images, but actually transforms them to monochrome.

A. Mapping to an Illumination-Invariant Chromacity Space

Recently, Ratnasingam and McGinnity [3] presented a method for mapping three image sensor responses (e.g., RGB colour space) to an illumination-invariant chromacity space, $\mathcal{I}$. The standard approach is to assume that the spectral sensitivity of the image sensor is infinitely narrow and that the spectrum of daylight can be approximated by a black body [3][7]. Under these assumptions, one can show that the output spectrum of a blackbody can be separated into three independent components: (i) a wavelength component, (ii) a reflectance component, and (iii) an illuminant component. For more details, the reader is referred to [7].

Using the illumination-invariant feature space from Ratnasingam and McGinnity [3], we can map the three colour channels in a raw image, $\{R_1, R_2, R_3\}$, to an illumination-invariant intensity, $\mathcal{I}$, according to

$$\mathcal{I} = \log(R_2) - \alpha\log(R_1) - \beta\log(R_3), \tag{1}$$

where $\{\alpha, \beta\}$ are channel coefficients, which are subject to the following constraint:

$$\frac{1}{\lambda_2} = \frac{\alpha}{\lambda_1} + \frac{\beta}{\lambda_3}, \tag{2}$$

with $\{\lambda_1, \lambda_2, \lambda_3\}$ being the peak sensitivity wavelengths for each image sensor[1]. See Figure 3 for some examples of this image transformation. Note, however, that this image transformation adds noise, which is particularly prominent in the foreground on the road. The significance of this will be discussed in more detail in the next section.

B. Combined Localisation System

Transforming the live image stream using (1) can be performed on a per-pixel basis, and is therefore inexpensive,

[1]These can be gathered from the sensor datasheet. In our experiments, we used $\alpha = 0.4800$ and $\beta = 0.5065$ for a Point Grey Bumblebee2 camera.

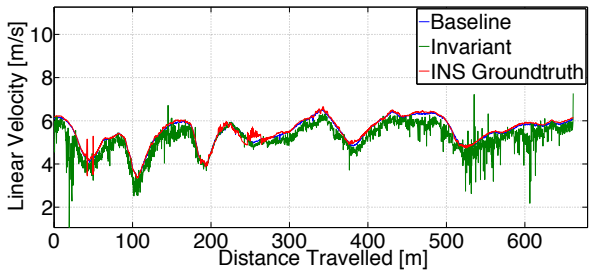


Fig. 5. Representative velocity estimates for a loop around our Begbroke Science Park. Note how the estimates using the lighting invariant images are noisier and appear to have a slight bias when compared to groundtruth.

allowing us to run this thread alongside the baseline system. Our strategy for the combined system is quite simple. We run both VO streams in parallel and when we are able to localise using the raw images (i.e., the baseline system), we take that estimate, otherwise, we switch to the lighting-invariant estimate. The reason we default to the baseline system is highlighted in Figure 5, which shows a representative velocity profile both with and without using the illumination-invariant image transform. There are two main differences that can be observed. The first is that the illumination-invariant estimates are noisier, which is likely due to the noise added by the pixel-wise transform. The second and more interesting difference is that there appears to be a slight bias in illumination-invariant estimates. We believe that this is a function of the feature distribution that results when using the illumination-invariant images. It appears that a lot of the high-frequency noise in the near field is amplified, meaning that fewer near-field features are detected. As a result, the feature distribution appears to be strongly biased towards the upper region, typically representing distant features. As there exists a known bias in stereo [18] (with a strong relationship to range), we believe this is the most likely explanation. Thus, owing to the increased noise and slight bias, fusing the estimates seemed suboptimal. Instead, we switch between the two system, with the policy of defaulting to the baseline system when possible. A block-flow diagram of our system is provided in Figure 4.

It is important to note that we can also perform the following trick for improved performance. Since the VO estimates using lighting invariant images are not as accurate as baseline system, we can use the live VO estimate from the baseline system to perform the feature prediction in the lighting-invariant feature space (as described in Section II-B). In otherwords, we can use the most recent frame-to-frame VO estimate from the baseline system to help inform the lighting-invariant VO pipeline where to look.

IV. EXPERIMENTS AND RESULTS

In this section, we present a series of localisation results both with and without the use of lighting-invariant imagery. We collected 15 visual surveys around the Begbroke Science Park with the focus on capturing more challenging lighting conditions. In Figure 6 we show some examples of the extreme visual variation encountered along parts of the route.

To clarify terminology, the system that does not use invariant imagery (RGB only) is the *baseline system*, the system that uses invariant imagery only is the *invariant system*, and the system that combines them is the *combined system*.

Fig. 6. Sample images gathered under a shadowy area in our Begbroke datasets. These areas prove to be very challenging for our baseline system due to the extreme variations in lighting.

For each of the 15 datasets, we used an exhaustive leave-one-out approach, whereby each dataset was taken as the live image stream, and localisation was performed against the remaining 14 datasets in turn.

TABLE I

COVERAGE RESULTS COMPARING OUR COMBINED SYSTEM VERSUS THE BASELINE SYSTEM. COVERAGE IS DEFINED AS THE NUMBER OF SUCCESSFULLY LOCALISED FRAMES AS A FRACTION OF THE TOTAL NUMBER OF CAPTURED FRAMES, AVERAGING OVER 14 TRAINING DATASET PER TEST DATASET.

Dataset Number	Baseline System	Combined System
1	79.93%	83.19%
2	92.68%	95.74%
3	91.12%	94.59%
4	95.81%	96.65%
5	94.19%	95.80%
6	93.64%	95.74%
7	95.64%	98.30%
8	96.29%	97.60%
9	94.75%	97.30%
10	93.90%	95.61%
11	83.47%	89.35%
12	95.88%	97.54%
13	91.87%	95.01%
14	86.58%	89.55%
15	97.33%	98.53%
Average	**92.17%**	**94.68%**

Table I presents the percentage coverage using each of the 15 datasets as the live run. We define percentage coverage as the number of successfully localised frames versus the total number of frames, averaged over the 14 datasets compared against. We found that our INS system was not reliable for groundtruthing due to significant GPS drift (on the order of meters). Instead, we took the approach of Churchill and

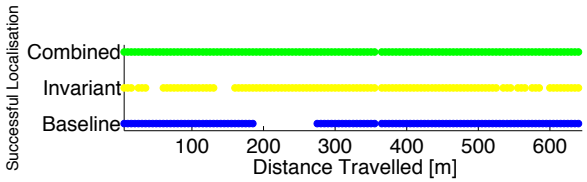


Fig. 7. One vs. One localisation result. The localisation performance of the three systems (i.e., the baseline, the illumination invariant system, and combined). Points indicate successful localisation. Between 190 m and 280 m invariant thread is able to localise where the baseline thread cannot. By taking the union of the two our combined system is more robust.

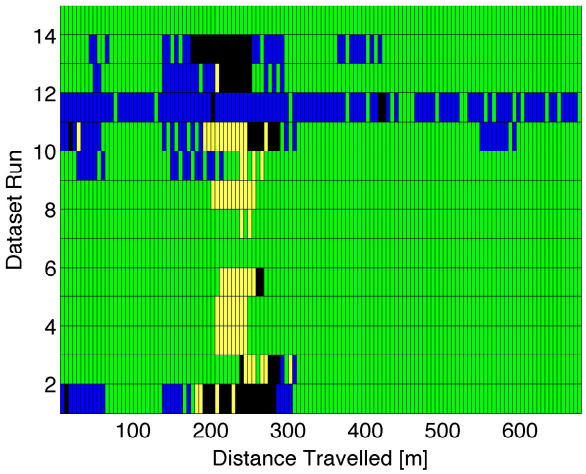


Fig. 8. One vs. All localisation results. The localisation characteristics of a single dataset (used as the live image stream) compared against the remaining 14 datasets. Each row corresponds to one of the 14 datasets, and the x-axis shows distance travelled. Blue indicates when only the baseline system localised, yellow indicates when only the invariant system localised, green is when both the baseline and invariant successfully localised, and black areas indicate localisation failures of both systems. By incorporating the invariant system we are able to localise successfully over a larger area.

Newman [2], which uses the localisation chain to predict the frame-to-frame motion and compares that with the VO estimate. If the two estimates disagree by a certain threshold then it is classified as a localisation failure.

In all cases the invariant system provides improvement to the baseline system, meaning the combined system *always* out-performs the baseline. An important result here is that our baseline system already performs well despite the difficult conditions. However, in the context of long-term autonomy for robotics, robustness is key, so any increase in reliability is important. We will show shortly that with the combined system we achieve significantly shorter distances travelling open loop during localisation failures.

Figure 7 shows the localisation performance of one live run versus one other dataset. In this figure, coloured points indicate a successful localisation for the specified system, while an absence of data represents a localisation failure. For this particular run, we see that the baseline system failed to localise over a 90 m section. However, because we have the invariant system running in parallel, which was able localise in this area, the combined system is able to localise for

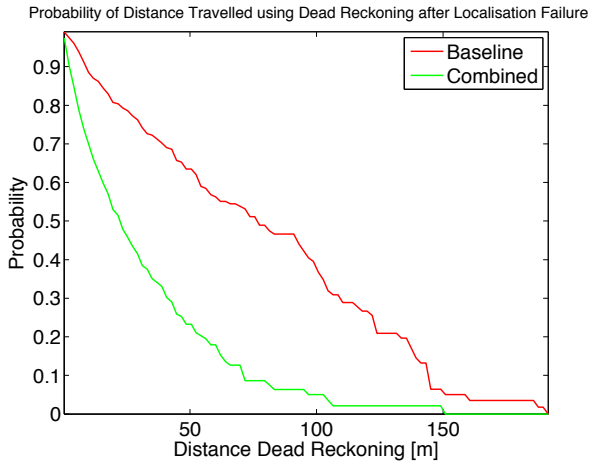


Fig. 9. Given a localisation failure, this plot shows how far the system is likely to travel before re-acquiring a localisation, i.e. how long it will have to travel using dead reckoning alone. In other words, this is $P(\text{dropout} = X)$, where X is distance traveled. We see that the combined system is likely to travel significantly shorter distances compared to the baseline after a localisation failure.

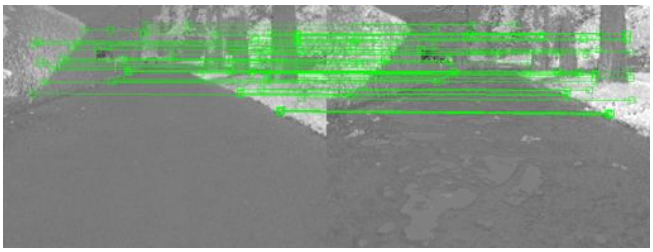
almost all of the route. Figure 10 show representative cases where the invariant localisation thread was successful while the baseline was not, and vice versa.

Figure 8 shows the performance of a single dataset used as the live image stream versus all 14 remaining datasets (along the y-axis). It is a graphical representation of one of the rows in Table I. In this plot, yellow indicates regions where only the invariant system could successfully localise. Here we see there is a region between 200-300 m along the route where the baseline thread repeatedly struggles, due to the challenging lighting variation (see Figure 6). It should also be noted that the invariant thread does not always contribute. The blue regions in Figure 8 indicates areas where only the baseline thread was successful. By taking the union of the two threads we have improved the robustness of our system.

We refer the reader to Figure 9, which is the key result of this paper. Given that this is a localisation system, the primary concern is exposure to extended periods of time or travel in which we fail to localise. During these periods we must fall back to deadreckoning from Visual Odometry—however good that is we are still effectively running "open loop". Figure 9 shows that the system we propose here, which leverages illumination-invariant colour spaces, a dual-processing pipeline, and a carefully informed search policy for feature associations, produces a performance far superior to the baseline system. For example, the likelihood of the system travelling blind for up to 100 m is close to 40% with the baseline system, whereas the with the combined system, the likelihood is just 5%.

V. CONCLUSION

Dealing with severe lighting changes is a critical requirement for long-term, persistent navigation in outdoor environments. We believe that the approach presented here will help move us in the right direction in order to tackle this



(a) Successful localisation under the trees. Data associations shown in green.

(b) Failed localisation under the trees. No successful matches.

(c) Failed localisation near a car park. No successful matches.

(d) Successful localisation near a car park. Data associations shown in green.

Fig. 10. Examples where the lighting-invariant images helped the system localise under a very shadowy region (top row) and where the lighting-invariant images failed to localise (bottom row). As can be seen, the image transform adds artefacts, which can sometimes result in fewer matches. However, the benefit of running this system becomes clear when looking at regions with high visual variability caused by external illuminates.

challenging problem. We have argued for the use of lighting invariant image transforms as a way to ease the difficulties arising from imaging in varying lighting conditions. We have shown that by folding this transform into an additional image processing pipeline, we can substantially reduce our exposure to having to deadreckon through long periods of localisation failure. The additional cost of our extension is a fixed price while the benefits are often substantial and never negative.

VI. ACKNOWLEDGMENTS

The authors wish to acknowledge the following funding sources. Colin McManus is supported by the Nissan Motor Company. Paul Newman and Winston Churchill are supported by EPSRC Leadership Fellowship Grant EP/J012017/1. The authors also wish to thank Peter Corke for setting our thinking about colour in motion in 2012.

REFERENCES

[1] C. McManus, W. Churchill, A. Napier, B. Davis, and N. P., "Distraction suppression for vision-based pose estimation at city scales," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, 6-10 May 2013.

[2] W. Churchill and P. Newman, "Practice makes perfect? managing and leveraging visual experiences for lifelong navigation," in *Proceedings of the International Conference on Robotics and Automation*, Saint Paul, Minnesota, USA, 14-18 May 2012.

[3] S. Ratnasingam and T. McGinnity, "Chromaticity space for illuminant invariant recognition," *IEEE Transactions on Image Processing*, vol. 21, 2012.

[4] M. Ebner, *Color Constancy*. Wiley-IS&T Series in Imaging Science and Technology, 2007.

[5] G. Finlayson, M. Drew, and C. Lu, "Intrinsic images by entropy minimization," *Computer Vision-ECCV 2004*, pp. 582–595, 2004.

[6] G. Finlayson, S. Hordley, C. Lu, and M. Drew, "On the removal of shadows from images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 1, pp. 59–68, 2006.

[7] S. Ratnasingam and S. Collins, "Study of the photodetector characteristics of a camera for color constancy in natural scenes," *Journal of the Optical Society of America A*, vol. 27, no. 2, pp. 286–294, Feb 2010. [Online]. Available: http://josaa.osa.org/abstract.cfm?URI=josaa-27-2-286

[8] R. Guo, Q. Dai, and D. Hoiem, "Single-image shadow detection and removal using paired regions," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2033–2040.

[9] J. Zhu, K. Samuel, S. Masood, and M. Tappen, "Learning to recognize shadows in monochromatic natural images," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 223–230.

[10] V. Kwatra, M. Han, and S. Dai, "Shadow removal for aerial imagery by information theoretic intrinsic image analysis," in *Computational Photography (ICCP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1–8.

[11] M. Milford and G. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Saint Paul, Minnesota, USA, 14-18 May 2012.

[12] P. Corke, R. Paul, W. Churchill, and P. Newman, "Dealing with Shadows: Capturing Intrinsic Scene Appearance for Image-based Outdoor Localisation," *IEEE International Conference on Intelligent Robots and Systems*, 2013.

[13] W. Maddern and S. Vidas, "Towards Robust Night and Day Place Recognition using Visible and Thermal Imaging," *Robotics Science and Systems*, 2012.

[14] C. McManus, P. Furgale, B. Stenning, and T. D. Barfoot, "Visual Teach and Repeat Using Appearance-Based Lidar," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[15] E. Rosten, G. Reitmayr, and T. Drummond, "Real-time video annotations for augmented reality," in *Advances in Visual Computing*, 2005.

[16] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2012.

[17] A. Davison, I. Reid, N. Motlon, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, 2007.

[18] G. Sibley, "Long range stereo data-fusion from moving platforms," Ph.D. dissertation, University of Southern California, 2007.