Lighting Invariant Urban Street Classification

Ben Upcroft[1], Colin McManus[2], Winston Churchill[2], Will Maddern[2] and Paul Newman[2]

***Abstract*— In this paper we propose the hybrid use of illuminant invariant and RGB images to perform image classification of urban scenes despite challenging variation in lighting conditions. Coping with lighting change (and the shadows thereby invoked) is a non-negotiable requirement for long term autonomy using vision. One aspect of this is the ability to reliably classify scene components in the presence of marked and often sudden changes in lighting. This is the focus of this paper.**

Posed with the task of classifying all parts in a scene from a full colour image, we propose that lighting invariant transforms can reduce the variability of the scene, resulting in a more reliable classification. We leverage the ideas of "data transfer" for classification, beginning with full colour images for obtaining candidate scene-level matches using global image descriptors. This is commonly followed by superpixel-level matching with local features. However, we show that if the RGB images are subjected to an illuminant invariant transform before computing the superpixel-level features, classification is significantly more robust to scene illumination effects.

The approach is evaluated using three datasets. The first being our own dataset and the second being the KITTI dataset using manually generated ground truth for quantitative analysis. We qualitatively evaluate the method on a third custom dataset over a 750 m trajectory.

I. INTRODUCTION

Vision-based classification methods heavily rely on the similarity in appearance between training data. However, lighting variation can cause the very same scene or object to appear very different from one image to the next. This difference can cause adverse effects to classification algorithms resulting in poor performance -"it just looks so different". For autonomous systems operating over extended time scales, such as days and weeks, illumination changes are detrimental if high level understanding from low level pixel data is to be achieved.

Over recent years, illuminant invariant transforms have been proposed which are able to significantly reduce this illumination variation in outdoor scenes [1], [2]. A number of these transforms reduce a three-channel RGB image to a single channel colour space. Therefore, the invariance provided by these transforms comes with the reduction in discriminative information available in the colour channels. As a result, classification solely dependent on lighting invariant transforms of an image can often fail.

Rather than using one or the other, we show that RGB images allow excellent context matching between images, while illuminant invariant images provide very good local discriminitave capabilities within an image. Data transfer methods for image classification enable the combination of these images by exploiting the two stage process of scene-level matching using global image features from the RGB images and local super-pixel feature matching from the lighting invariant images. By employing the two types of data as input to the classification problem, the advantages of both are gained, resulting in robust classification even under highly varying lighting conditions (Fig. 1).

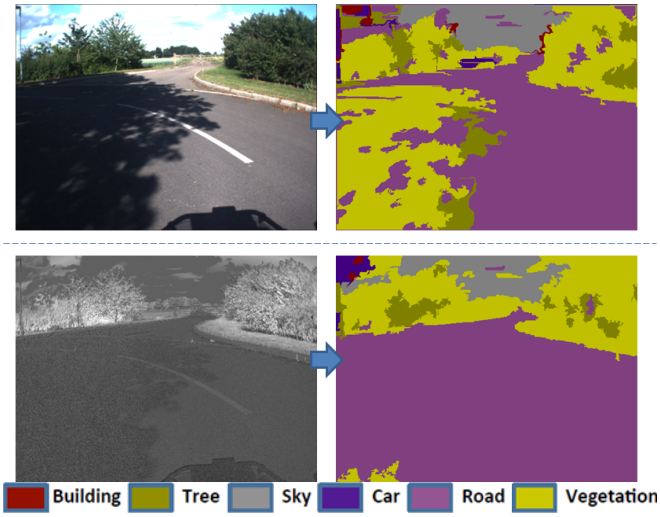

Fig. 1. A comparison of classification results for the same scene using an RGB image (top row) and a lighting invariant image (bottom row). Top row: RGB image and its corresponding classification. Bottom row: lighting invariant version of the same image and its corresponding classification. Note how the shadow on the road is correctly classified when using the lighting invariant image.

This paper is organised as follows. Section II will discuss related work in scene classification techniques and the use of lighting invariant images in robotics applications. Section III provides a system overview, summarising the basic segmentation algorithm, our lighting-invariant image transform, and describing how we incorporate the lighting-invariant images in our pipeline. Section IV presents quantitative results on two challenging outdoor datasets: (i) a custom dataset collected at Oxford, and (ii) the KITTI dataset. A qualitative analysis on over 750 m of data around the Oxford Begbroke Science Park is also presented. Section V provides a discussion of our method and results, with an emphasis on the current limitations and how we plan on moving forward to enable robust scene classification in the presence of extreme illumination changes.

[1]Ben Upcroft is with the School of Electrical Engineering and Computer Science, Queensland University of Technology, Australia. e-mail:ben.upcroft@qut.edu.au.

[2]Colin McManus, Winston Churchill, Will Maddern, and Paul Newman are with the Mobile Robotics Group, University of Oxford, UK. e-mail: colin,winston,wm,pnewman@robots.ox.ac.uk.

II. RELATED WORK

Semantic multi-class segmentation has seen consistent activity in the computer vision domain [3], [4], [5], [6]. These approaches commonly employ offline training or learning using a variety of feature representations like appearance, colour, shape or depth to model each class in the scene. These features are then combined to build a spatial smoothness prior for a Markov Random Field (MRF) or Conditional Random Field (CRF) optimisation to infer a label for each pixel or superpixel (i.e., pixel sets with homogeneous attributes). These parametric approaches can take hours to days to train, where training needs to be repeated if new example classes are included in the dataset.

To complement the parametric techniques, data-driven approaches for scene inference have been investigated [7], [8], [9], [10], [11]. These approaches do not require a training step and examples of new classes can be simply added to the dataset. Initial work in this area concentrated on global feature descriptors to match entire images [12], [9] but have more recently included a second step in which local features within an image are also considered [7], [8], [10], [11].

This paper focuses on classification in urban street scenes which has increased in interest with autonomous motor vehicles requiring persistent operation in a large variety of environments and lighting conditions [13]. Bileschi *et al.* and Geiger *et al.* [14], [15] investigated urban scene classification but focussed on a very limited set of classes. Geometry constraints such as ground planes and specification of horizontal and vertical surfaces have enabled large improvements in classification such as in [16], [10].

Temporal coherence across images was implemented by Tighe and Lazebnik demonstrating impressive results over a very large dataset [11]. However, under significant lighting variation, classification failures tend to increase. Another method employing temporal coherence was demonstrated by Sengupta *et al.* where multiple views were fused together to form a global overhead image [17]. Our work would benefit from both these techniques but the focus of this paper is to improve individual frame classification before any temporal coherence is enforced.

Another form of employing consistency across frames is to use geometrical coherence [18], [19]. Sengupta *et al.* and He *et al.* projected semantic labels into a 3D point cloud computed from stereo geometry [18], [19]. Enforcing 3D consistency across semantic labels allows for increased robustness from individual frame classification noise but can still suffer from varying lighting conditions.

In all the above cases, large lighting variation, such as shadows, reduce the reliability of classification from image to image. The computer vision community have explored shadow detection and removal in images by learning classifiers based on intensity and colour cues [20], [21]. Within robotics, Park and Lim [22] detected shadows cast on the road by vehicles and surrounding structure but do not consider shadows from natural objects in the scene. Corke *et al.* investigated localisation using Finlayson *et al.*'s lighting invariant transform and showed preliminary results for classification robustness [23], [24]. Our work is inspired by this research where we focus on increasing the robustness of classification by introducing a similar lighting invariant transform as an input to our classification pipeline.

III. SEMANTIC SEGMENTATION

We adopt the classification method of Tighe and Lazebnik [10] in which RGB images are used for obtaining candidate scene-level matches using global image descriptors. Superpixel-level matching between each of the candidate images is subsequently performed. Class labels are then transferred from the most closely matched superpixel providing a dense classification for each image. These labels are refined using Markov Random Field optimisation to take into account the pixel-wise neighbourhood information. We give some details of Tighe and Lazebnik's method here for completeness and to highlight how our method takes advantage of the separate global descriptor and local superpixel matching steps.

A. Scene-Level Image Matching

Data-driven methods commonly begin by obtaining a small set of candidate images (known as the "retrieval set") from a database that most closely match the current query image. This ensures that scene-level context is maintained for subsequent local superpixel matching. To find a retrieval set that best represents the query image, we use three types of global image features: 1. spatial pyramid [25]; 2. gist [26], and 3. colour histogram. Each feature is of the same dimensionality as in [10].

The database images are then ranked according to their similarity score for each feature. The highest rank from the three scores is then assigned to that image. In the work presented in this paper, we take a relatively small set (between 5 and 30) of the top-ranked images from a database of 40 (See Fig. 2 for an example retrieval set). The size of the retrieval set and database is significantly less than what is normally considered useful in data transfer problems. However, we have previously shown successful results using these smaller datasets [19] and posit that due to the high similarity of the data (relative to the image diversity seen in most data transfer applications), the number of images is sufficient for the problem. As more diversity is required, the database can be updated accordingly without the need to relearn a parametric model.

B. Superpixel Feature Matching

Superpixels are computed using mean shift [27] and then a variety of descriptors are used to represent each superpixel, including relative position of the superpixels in an image, SIFT [28], Texton histograms [29], colour histograms, and gist [26]. Features for each of the superpixels in the training database are computed and stored with the class label (obtained from hand-labelling).

Fig. 2. An example of a typical query image from the Oxford dataset (left image). Right images: The three top ranked images in the retrieval set for the left query image. Note the similarity between the images which ensures that context is maintained for the lower level superpixel feature matching.

1) Lighting Invariance: We divert from Tighe and Lazebnik's method at this point, not in the matching process itself but in the data used to do the matching. Rather than using the original RGB images to compute superpixel features, we substitute the images for lighting invariant versions. Lighting variation in RGB images can cause incorrect classifications for the following reasons. Firstly, the superpixel generation will incorrectly segment parts of the same object because of shadows (recall that superixels are generated using the mean shift algorithm). Secondly, because the colour of shadowy regions is very different, the local descriptors can end up far apart in feature space. Lighting invariant transforms can mitigate some of these issues, by providing a more accurate segmentation based on the *true colour* of the scene (see Fig. 1 for example).

Methods that exploit the log-ratio of photodetector responses have the advantage of being simple but also applicable to a variety of lighting conditions, including sunlight [30], [24]. We follow the approach in [2] and use a one-dimensional feature space F consisting of three linear sensor responses, $\{R, G, B\}$, corresponding to peak sensitivities at wavelengths $\{\lambda_R, \lambda_G, \lambda_B\}$:

$$F = \log(G) - \alpha \log(B) - (1-\alpha)\log(R) \tag{1}$$

If the illumination source can be approximated by a black-body radiator, the feature space in Eq. 1 minimises the variance caused by scene geometry, illuminant intensity and source spectrum, provided the parameter α satisfies the following constraint [2]:

$$\frac{1}{\lambda_G} = \frac{\alpha}{\lambda_B} - \frac{(1-\alpha)}{\lambda_R} \tag{2}$$

The values for $\{\lambda_R, \lambda_G, \lambda_B\}$ are set to the peak spectral response for each colour channel of the Bayer filter from the datasheet of the image sensor.

Fig. 3. Example lighting invariant images from the Oxford dataset. Shadows become less dominant in the images and the saliency of the material properties in the scene are increased.

An example of the lighting invariant transform is shown in Fig. 1 and 3. Note that the shadow effects are significantly reduced. Although this method is able to reduce the effect of shadows, it can be quite sensitive if a single RGB channel is over-exposed or under-exposed as can be seen in Eq. 1. This can result in noisy images after the transform especially if compression artifacts are present.

Note that while we have found lighting invariant images to have increased benefit at the superpixel matching stage, they do not perform as well as RGB at the scene-level matching stage. We believe that this is due to the colour histograms providing a rich context-related feature lacking in the grey-scale histograms from lighting invariant images.

Inclusion of the lighting invariant transform into the classification will be referred to as the LI method while the baseline method of Tighe and Lazebnik's will be referred to as the RGB method throughout the rest of the paper.

C. Label Transfer

Once all the superpixel features have been computed from the illumination-invariant images, we choose class c for superpixel s_i if:

$$P(c|s_i) \geq P(\bar{c}|s_i) \tag{3}$$

where $\bar{c}$ is the set of classes not including c.

To determine the class we can use Bayes Rule and get:

$$\frac{P(s_i|c)P(c)}{P(s_i)} \geq \frac{P(s_i|\bar{c})P(\bar{c})}{P(s_i)} \tag{4}$$

After rearranging, a likelihood ratio score $L(s_i, c)$ for each test superpixel and each class in the retrieval set can be obtained:

$$\begin{aligned} \frac{P(s_i|c)}{P(s_i|\bar{c})} &\geq \frac{P(c)}{P(\bar{c})} \\ L(s_i, c) &\geq \frac{P(c)}{P(\bar{c})} \end{aligned} \tag{5}$$

The likelihood ratio can then be computed as the normalised distance between features of each superpixel in the query image and the nearest neighbours of the superpixels in the retrieval set. For further details, see Tighe and Lazebnik [10]. A label for each superpixel can now be assigned by maximising this likelihood ratio score.

Finally, this labelling is used to initialise an optimisiation of a standard MRF energy function as in [10], [19].

IV. EXPERIMENTS

To demonstrate the proposed system, we used the publicly available KITTI dataset [31] and two custom datasets collected in Oxford over multiple kilometres and at varying times of the day. The KITTI and one of the Oxford datasets (with 41 and 40 database images respectively, and 10 test images for each) were used for quantitative analysis. The test images were hand-labelled to provide a ground truth comparison[1]. We also qualitatively evaluate the proposed method on the second Oxford dataset over a 750 m trajectory. For all KITTI and Oxford images we set $\alpha = 0.48$. Note that we attempted to use the extensive CamVid dataset for further evaluation but the sensitivity of the lighting invariant transform to the compression artifacts present in the dataset resulted in extremely noisy images.

Our implementation in Matlab (a modification of the publicly available code provided by [10]), takes approximately 30 s to classify a query image once features from the database have been pre-computed. Most processing time can be attributed to feature extraction and matching which can be significantly improved with a GPU implementation.

We directly compare the hybrid RGB and Lighting Invariant method with Tighe and Lazebnik's state of the art method [10], [11] but do not include geometric or temporal constraints. The two methods are labelled "LI" and "RGB" respectively, in the following bar graphs.

A. Quantitative Evaluation

For semantic accuracy evaluation, we use the evaluation measures similar to those defined in [18] to compute per-class Recall (classification rate) (R), Average Recall (AR), Global Recall (GR), and $F1$ score defined below:

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}} \tag{6}$$

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}} \tag{7}$$

[1] Our hand-labelled images for the KITTI dataset are available at `http://wiki.qut.edu.au/display/cyphy/Datasets`.

Fig. 4. Noise introduced by the lighting invariant transform in the KITTI dataset. The wide field of view results in many pixels with a single dominant RGB channel caused by the Bayer encoding across neighbouring pixels. As can be seen in Eq. 1, relatively small or large values in any channel will be amplified in the resultant transform.

$$AR = \frac{1}{|c|} \sum_c R_c \tag{8}$$

$$GR = \frac{\sum_c N_{tp}}{\sum_c (N_{tp} + N_{fn})} \tag{9}$$

$$F1 = 2\frac{P \times R}{P + R} \tag{10}$$

where $|c|$ is the total number of classes, R_c is the recall for class c, N_{tp}, N_{fp}, and N_{fn} refer to the number of true positive, false positive, and false negative pixels respectively. The Global Recall evaluates the overall ratio of correct labelling, the Average Recall evaluates the average per-class recall score, and the F1 score is a measure of the per class accuracy.

1) KITTI Dataset: We evaluate our method on images obtained from the KITTI datasets [31]. The images include common objects such as pedestrians, bicyclists, cars, trees, and buildings at a resolution of 1242×375. We manually annotated 51 images from different KITTI datasets to ensure a diverse set of scenes. We label the scene into 7 semantic classes, *i.e.*, Building, Car, Sky, Tree, Sidewalk, Road, VegetationMisc. These datasets are quite challenging, and even objects of the same class in the scene have different appearance.

One of the main drawbacks of the KITTI images are the artifacts introduced through Bayer encoding over a very wide field of view. The lighting invariant transform is severely effected by these artifacts resulting in noisy greyscale images (see Fig. 4). Despite the noisy lighting invariant transform, classification remained robust to shadows while RGB images alone suffered from incorrect classification. In particular, areas with significant shadows were greatly effected as can be seen in Fig. 5.

Fig. 6 compares the F1 score (top) and recall (bottom) for each of the classes, illustrating the robustness of the lighting invariance transform for superpixel matching. The only significant failure is with sidewalks and as can be seen in Fig. 4. The noise introduced by the artifacts in the KITTI

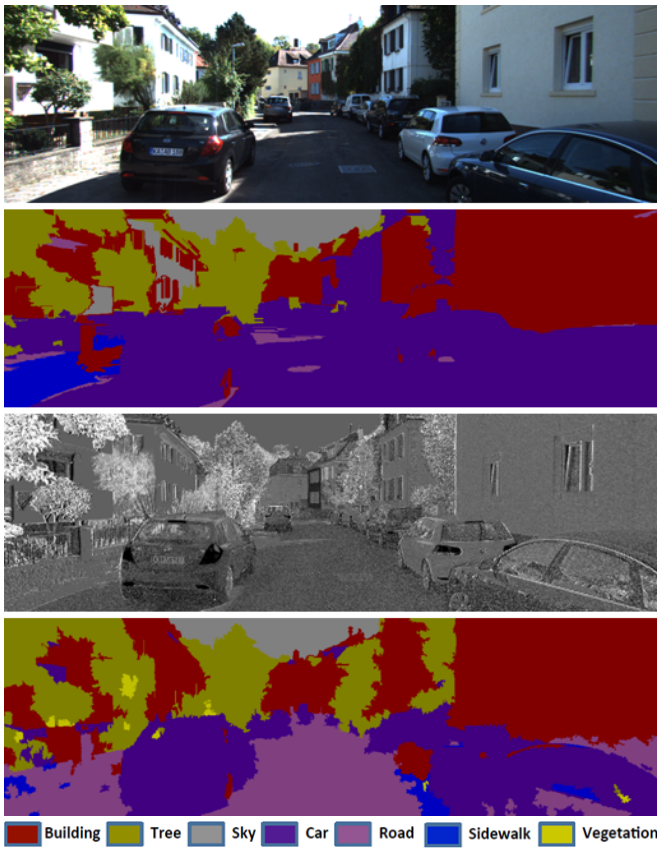

Fig. 5. Comparison between RGB (top two images) and LI (bottom two images) classification in the presence of significant shadows. The LI method is significantly more robust to these types of scenarios than the RGB method.

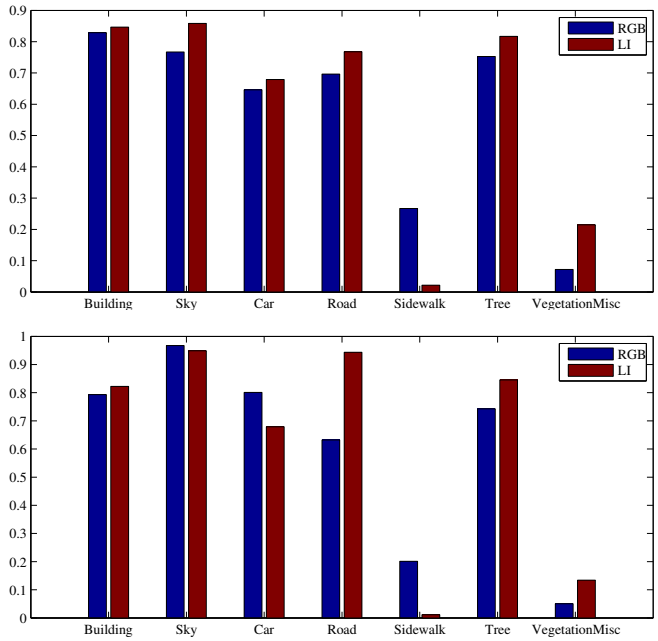

Fig. 6. Top plot: F1 score for the KITTI dataset. Bottom plot: Recall for the KITTI dataset. Blue and red bars refer to the per-class F1 score and Recall for the RGB and LI method respectively. The LI method, generally outperforms the baseline system in all categories, except sidewalks and cars. This is likely due to the noise introduced in the LI transformation, which blurs the boundaries between road and sidewalk and also can exhibit high noise around high reflectivity from the metal surfaces of a car.

Fig. 7. Example images from the Oxford dataset. Shadows from overhanging and nearby foliage is a major cause of failure for the baseline RGB classification method.

dataset blur the spatial and feature-space boundaries between sidewalks and roads. It can also be seen that the LI Recall for cars does not perform as well as the RGB method. This is most likely due to the noise introduced by high reflectivity of the cars and the resultant saturated pixels. Note though that the LI F1 score for cars slightly outperforms that of RGB where this score also includes Precision in the evaluation.

2) Oxford Dataset: The second quantitative analysis is performed on the custom Oxford dataset. The images contain urban scenes with significantly more vegetation than exhibited in the KITTI images. The resolution of the images are 512×384. Example images from the dataset are shown in Fig. 7. Due to the narrower field of view, the lighting invariant transform results in less noise than the KITTI images (Fig. 1) and as a result, a cleaner classification.

Figure 8 illustrates significant improvements for the F1 score (top) and recall (bottom) across most of the classes than that of RGB alone. Note that the building class has very low recall in both instances due to the small number of buildings in the database.

Table I shows the average/global recall for the two datasets. The addition of lighting invariant transforms at the local level outperform or equal RGB alone.

We wish to emphasise that although the improvements may seem modest, this is due to the fact that not all images exhibit large lighting variation. As we are interested in developing techniques for autonomous road vehicles, handling these edge cases (e.g., misclassification due to extreme lighting changes) must be addressed.

3) 750 m Oxford Dataset: Figure 9 shows a few examples from the 750 m dataset[2]. There were a number of cases where shadows severely effected the baseline RGB classification method whereas the LI classification proved more reliable in these areas. The bottom row shows a failure case for both the RGB and LI methods and we attribute this to the lack of training examples with signs in the scene. We also have noticed failure cases for the LI method when pixels are

[2]The accompanying video qualitatively compares the classification techniques over the entire sequence of images for the trajectory

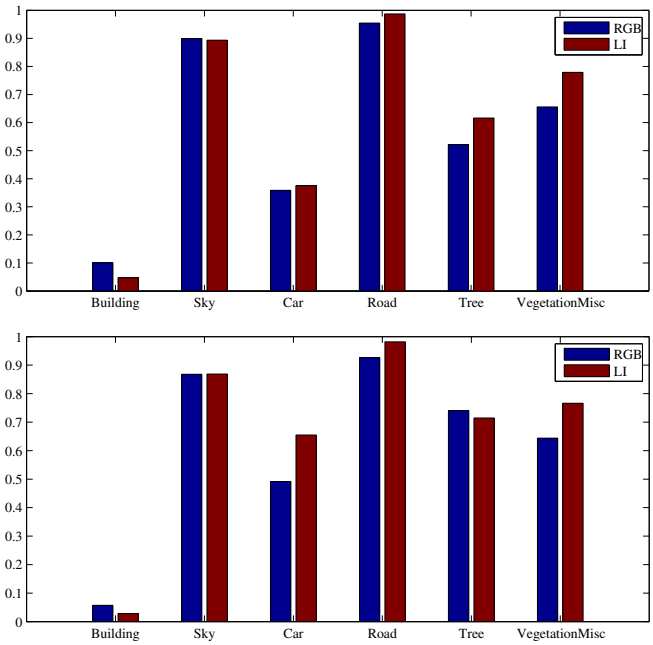

Fig. 8. Top plot: F1 score for the Oxford dataset. Bottom plot: Recall for the Oxford dataset. Blue and red bars refer to the per-class F1 score and Recall for the RGB and LI method respectively. The LI method, generally outperforms the baseline system in all categories, except buildings where both methods performed poorly. This is most likely due to the lack of buildings in the database.

TABLE I

GLOBAL AND AVERAGE RECALL FOR THE RGB AND LI METHODS.

Method	KITTI	Oxford
Global Recall		
RGB	0.60	0.62
LI	0.63	0.67
Average Recall		
RGB	0.60	0.53
LI	0.61	0.53

saturated in one of the colour channels resulting in significant texture change in the LI transform.

As classification methods push accuracy higher, it is sometimes difficult to evaluate when only 1-2% improvement is observed overall. It is therefore important to consider gross individual failures which would be detrimental to a persistent autonomous system. Two of these cases are highlighted in the top two rows of Figure 9. Although these are only "corner-cases", they significantly reduce the mean time to failure for a robotic vehicle.

These results point to the advantage of using alternate inputs to classification routines while ensuring we maintain the advantages of the the RGB methods developed throughout the community.

V. DISCUSSION

The use of lighting invariant transforms in classification problems shows much promise in increasing reliability under heavy shadows as shown in the previous results. However, the transforms can exhibit high sensitivity to saturation of

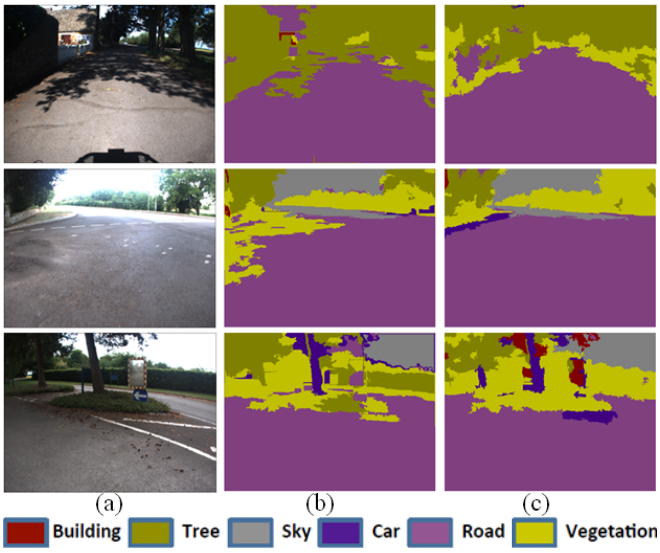

Fig. 9. (a) Sample images from the qualitative evaluation. Classification results for the corresponding sample images using the (b) RGB and (c) LI methods. The top two rows show examples where shadows severely effect the RGB method but the LI method copes well. The bottom row illustrates a failure case for both methods and is most likely due to the lack of similar images in the database.

images and compression artifacts (as experienced with the KITTI and CamVid datasets). Furthermore, objects in the scene with similar material properties such as the sidewalk, road, and even some buildings, can cause difficulties. The particular lighting-invariant transform used in this paper does not account for non-blackbody light sources (e.g., lights from other cars) and thus would not be suitable when exposed to these types of sources.

However, we demonstrate in this paper that in a number of very difficult scenarios, lighting-invariant images can be extremely robust. We anticipate that rather than choosing between an RGB or LI approach, that the fusion of the two would prove beneficial. Another promising approach would be to augment RGB images with the lighting invariant transform to produce a four-channel image and proceed with the method outlined in this paper.

Further improvements in this technique would be to increase the dataset and retrieval set sizes. As demonstrated in the computer vision community, dataset size has a strong role to play. Finally, temporal coherence would dramatically improve results as demonstrated by Tighe and Lazebnik using RGB alone [11]. As discussed though, any improvement at the single image stage will always ensure increased reliability in any extended pipeline.

VI. CONCLUSION

We present a simple extension to a state-of-the-art scene classification framework [10], [11] to improve robustness to extreme lighting variation. To our knowledge, this is the first comprehensive application of lighting-invariant transforms to street-scene classification. Our method outperformed the baseline in almost all categories on two different urban datasets. To enable long term autonomy, we believe that

this is one step in the direction to increasing reliability for scene classification in outdoor environments over extended timescales.

ACKNOWLEDGMENT

Ben Upcroft was supported by Australian Research Council project DP110103006 Lifelong Robotic Navigation using Visual Perception. Paul Newman and Winston Churchill were supported by an EPSRC Leadership Fellowship, EPSRC Grant EP/I005021/1. Authors thank Hu He for Matlab code to help convert between formats and for evaluation. We thank Alex Stewart and Hugo Grimmett for valuable suggestions on this paper. The authors would also like to thank Peter Corke for motivating this line of research in 2012.

REFERENCES

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

[2] 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*, vol. 27, 2010.

[3] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr, "Associative hierarchical crfs for object class image segmentation," in *International Conference in Computer Vision*, 2009.

[4] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *European Conference on Vision Conference*, 2008.

[5] C. Zhang, L. Wang, , and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," in *European Conference on Computer Vision*, 2010.

[6] P. Sturgess, K. Alahari, L. Ladicky, , and P. Torr, "Combining appearance and structure from motion features for road scene understanding," in *British Machine Vision Conference*, 2009.

[7] C. Liu, J. Yuen, A. Torralba, J. Sivic, , and W. Freeman, "Sift flow: Dense correspondence across different scenes," in *European Conference on Computer Vision*, 2008.

[8] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, p. 2368, 2011.

[9] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition."

[10] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *European Conference on Computer Vision*, 2010.

[11] ——, "Superparsing: Scalable nonparametric image parsing with superpixels," in *International Journal of Computer Vision*, 2012.

[12] J. Hays and A. A. Efros.

[13] W. Churchill and P. Newman, "Continually improving large scale long term visual navigation of a vehicle in dynamic urban environments," in *Proc. IEEE Intelligent Transportation Systems Conference (ITSC)*, September 2012.

[14] S. Bileschi and L. Wolf, "A unifed system for object detection, texture recognition, and context analysis based on the standard model feature set," in *British Machine Vision Conference*, 2005.

[15] A. Geiger, M. Lauer, and R. Urtasun, "A generative model for 3d urban scene understanding from movable platforms," in *International Conference on Computer Vision and Pattern Recognition*, 2011.

[16] I. Posner, M. Cummins, and P. Newman, "A generative framework for fast urban labeling using spatial and temporal context," *Journal of Autonomous Robots*, vol. 26, p. 153.

[17] S. Sengupta, P. Sturgess, and L. L. P. Torr, "Automatic dense visual semantic mapping from street-level imagery," in *International Conference on Intelligent Robots and Systems*, 2012.

[18] S. Sengupta, E. Greveson, A. Shahrokni, and P. Torr, "Urban 3d semantic modelling using stereo vision," in *International Conference on Robotics and Automation*, 2013.

[19] H. He and B. Upcroft, "Nonparametric semantic segmentation for 3d street scenes understanding," in *International Conference on Intelligent Robots and Systems*, 2013.

[20] J. Zhu, K. Samuel, S. Masood, and M. Tappen, "Learning to recognize shadows in monochromatic natural images," in *International Conference on Computer Vision and Pattern Recognition*, 2010.

[21] R. Guo, Q. Dai, and D. Hoiem, "Single-image shadow detection and removal using paired regions," in *International Conference on Computer Vision and Pattern Recognition*, 2011.

[22] S. Park and S. Lim, "Fast shadow detection for urban autonomous driving applications," in *International Conference on Intelligent Robots and Systems*, 2009.

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

[24] P. Corke, R. Paul, W. Churchill, and P. Newman, "Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation," in *International Conference on Intelligent Robots and Systems*, 2013.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *International Conference on Computer Vision and Pattern Recognition*, 2006.

[26] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Visual Perception, Progress in Brain Research*, vol. 155, 2006.

[27] L. H. K. Fukunaga, "The estimation of the gradient of a density function, with applications in pattern recognition," vol. 21, no. 1, p. 32.

[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[29] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *International Journal of Computer Vision*, vol. 43, no. 1, p. 7, 2001.

[30] G. D. Finlayson and M. S. Drew, "4-sensor camera calibration for image representation invariant to shading, shadows, lighting, and specularities," in *IEEE International Conference on Computer Vision*, 2001.

[31] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.