Exploiting 3D Semantic Scene Priors for Online Traffic Light Interpretation

Dan Barnes, Will Maddern and Ingmar Posner

Abstract— **In this paper we present a probabilistic framework for increasing online object detection performance when given a semantic 3D scene prior, which we apply to the task of traffic light detection for autonomous vehicles. Previous approaches to traffic light detection on autonomous vehicles have involved either precise knowledge of the relative 3D positions of the vehicle and the traffic light (requiring accurate and expensive mapping and localisation systems), or a classifier-based approach that searches for traffic lights in images (increasing the chance of false detections by searching all possible locations for traffic lights). We combine both approaches by explicitly incorporating both prior map and localisation uncertainty into a classifier-based object detection framework, generating a scale-space search region that only evaluates parts of the image likely to contain traffic lights, and weighting object detection scores by both the classifier score and the 3D occurrence prior distribution. We present results comparing a range of low- and high-cost localisation systems using over 30 km of data collected on an autonomous vehicle platform, demonstrating up to a 40% improvement in detection precision over no prior information and 15% improvement on unweighted detection scores. We demonstrate a 10x reduction in computation time compared to a naïve whole-image classification approach by considering only locations and scales in the image within a confidence bound of the predicted traffic light location. In addition to improvements in detection accuracy, our approach reduces computation time and enables the use of lower cost localisation sensors for reliable and cost-effective object detection.**

I. Introduction

Detection and interpretation of traffic lights is a crucial task for autonomous vehicles as well as an increasingly viable safety feature for advanced driver assistance systems (ADAS). The implementation of a system capable of reliably detecting and interpreting traffic lights in all conditions, such as darkness, rain and fog, remains a challenge for the computer vision and robotics community [1]–[4]. Although traffic lights which broadcast their state to vehicles and intelligent intersection traffic management have been proposed [5], such technology is years from widespread deployment and cannot be relied upon in the immediate future.

Approaches for fully-autonomous vehicles such as [1], [2] use an accurate 3D map of the environment with labelled traffic lights, along with high-accuracy localisation systems costing upwards of £100,000 [6]. Online detection and interpretation of a traffic light becomes a comparatively simple task of reprojecting the expected 3D location of the traffic light into the image and examining the relevant pixel values to determine the traffic light state. By constraining other aspects of the system such as the camera exposure value, these systems provide a computationally inexpensive method of traffic light state interpretation, but relies entirely on a high-accuracy and high-cost localisation system.

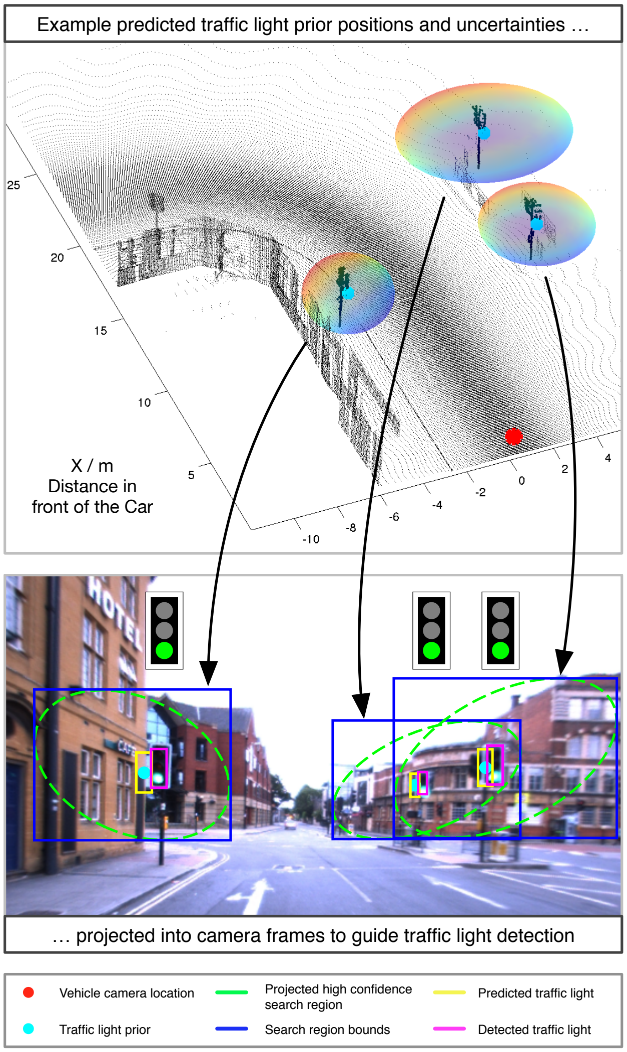

Example predicted traffic light prior positions and uncertainties …
25
20
15
10
5
X / m
Distance in
front of the Car
-10
-8
-6
-4
-2
0
2
4
HOTEL
… projected into camera frames to guide traffic light detection
● Vehicle camera location
● Traffic light prior
Projected high confidence search region
Search region bounds
Predicted traffic light
Detected traffic light

Fig. 1. Previously labelled traffic light prior locations and uncertainties are calculated as requested for online detection relative to the vehicle camera (top) based on localisation data. High confidence search regions are projected into the camera image, guiding the traffic light search in both scale and pixel location. Detection scores are weighted by the 3D occurrence prior, favouring search locations closer to the predicted traffic light location. Finally we calculate the most likely traffic light state.

In contrast, classifier-based approaches to object detection do not require prior map or localisation information, and are therefore attractive for low-cost traffic light interpretation systems. However, the lack of prior information about traffic light locations necessitates considering the entire image (both in position and scale) for possible traffic locations, effectively maximising the space in which to detect a false positive.

Authors are with the Mobile Robotics Group, University of Oxford, UK. daniel.barnes@keble.ox.ac.uk, {wm,ingmar}@robots.ox.ac.uk

In this paper we propose an object detection framework using a 3D semantic prior map to constrain online search regions, both in an attempt to provide efficient high performance detections (improving precision even when supplied with low accuracy localisation) and to reduce a combination of compute time and sensor cost. We focus specifically on traffic light detection and interpretation, yet there is no underlying obstacle to prevent our approach being applied to other detection targets. First we create a prior 3D map of the environment, including manually labelled locations of traffic lights, with a dedicated survey vehicle. At run-time we localise within the map using a range of navigation systems (consumer GPS, high-accuracy INS, stereo vision). We demonstrate that even minimal prior information about traffic light locations, such as the expected number of traffic lights in an image, can significantly increase detector performance compared to the baseline approach, and that performance uniformly improves with increasingly accurate (and increasingly expensive) localisation systems. By explicitly incorporating the positional uncertainty of both the prior map and localisation solution into the detection framework, we can simultaneously improve detector performance while reducing both computational requirements and hardware cost.

II. Related Work

In recent years, multiple approaches for traffic light detection and parsing for autonomous vehicles have been proposed and demonstrated in real-world conditions by numerous research groups [1]–[4], [7]. Both [1] and [2] require detailed prior 3D maps with labelled traffic lights, along with extremely accurate and high cost localisation systems to determine the location of the vehicle within the 3D map. Traffic light priors are projected into images as a predicted location to guide a search. As the localisation solution is so accurate, prior uncertainties are not used to explicitly weight detection scores [2] nor determine search region size [1].

There has been a considerable research effort on reliable unguided object detection in images, in particular that of pedestrians and traffic signs. Histogram of Oriented Gradients (HOG) descriptors [8] have been used to identify pedestrians and other objects with high precision and recall and have been used extensively in recent years. Newer methods such as discriminatively trained deformable parts models [9] and integral channel features [10] have also shown benefits such as increased robustness to occlusions and viewpoint changes. Although these detectors rely on no prior information, by not intelligently constraining the search region they are more susceptible to noise and false detections away from the true location.

Classification of colour dependant objects have relied upon fixing camera parameters to allow consistency between frames. Simple thresholds or histogram comparisons using colour spaces with separate illumination channels, such as Hue Saturation Variance (HSV) and LAB [2], [11] demonstrate high performance. Intelligent exploitation of additional constraints such as the aspect ratio of arrow versus circle bulbs [1] and prior colour knowledge of traffic lights can increase interpretation speed and precision further.

III. Semantic Prior Map

As with other approaches mentioned in Section II, we utilise prior information to improve traffic light detection. However, by capturing uncertainty in the object detection framework, we do not require a highly accurate prior map - we only require that the *uncertainty* in the prior map is recorded (e.g. as part of a SLAM framework). In our case the prior map consists of 3D pointcloud representation of the routes as well as manually tagged traffic light locations. When localised within the map, we are able to predict 3D traffic light positions relative to the vehicle and therefore project the traffic light locations onto the 2D image plane of a vehicle-mounted camera.

The prior map pointclouds are created from 2D laser and stereo camera data mounted to a survey vehicle and collected during normal driving. Relative transforms and uncertainties are calculated between successive stereo camera frames using the approach presented in [12]. 3D pointclouds are then generated by projecting 2D laser scans along the trajectory calculated by the stereo camera solution.

Each encountered traffic light is manually labelled in the 3D pointcloud at the centre of the amber bulb, as part of a post-processing stage requiring minimal labelling effort per traffic light. Traffic light priors are therefore uniquely identifiable by their laser measurement timestamp which, when localised relative to the prior map, are used to predict traffic light visibility and size in the camera view. An example 3D prior map is shown in Fig. 1, where a local pointcloud and labelled traffic light priors are projected in front of the vehicle, predicting their position in the camera image.

IV. Traffic Light Detection

In this section we present a traffic light detection framework which constrains the visual search region using the uncertainty of 3D priors. The goal is to both improve detection performance and computational efficiency, especially when faced with low accuracy online localisation.

A predicted traffic light location x^* in a camera image is defined in eq. 1, where c is the traffic light class and $x = (u, v, s)$ where (u, v) is the traffic light location on the image plane and s is the image pyramid scale.

$$x^* = \arg\max_x \; p(x|\, c) \tag{1}$$

The maximisation function in eq. 1 is derived from the likelihood of a traffic light in image patch x, $p(c|\, x)$, and traffic light occurrence prior distribution, $p(x)$. We treat the discriminative classifier output, $p(c|\, x)$, as likelihood in order to arrive at a posterior distribution over locations in the image, $p(x|\, c)$.

$$p(x|\, c) \propto p(c|\, x)\; p(x) \tag{2}$$

Implied here is the assumption that $p(c|\, x)$ is proportional to the support vector machine (SVM) [13] classifier score using Platt scaling [14] calculated from the HOG descriptor $\mathbf{H}$ evaluated from image patch x:

$$p(c|\, x) \propto \mathbf{w}^T\; \mathbf{H}|_x + b \tag{3}$$

The prior distribution $p(x)$, represented in Fig. 3 as the 3D Gaussian distribution $\mathbf{T}, \sum$, is therefore a function of vehicle position V, prior map M, traffic light position T and camera matrix K.

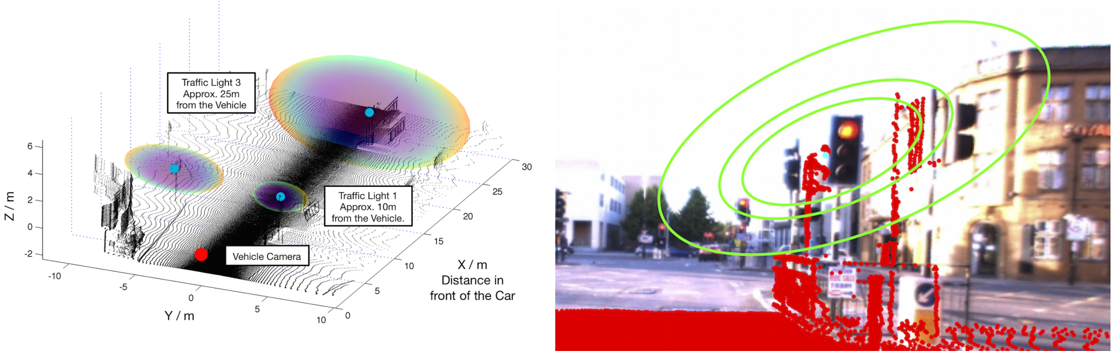


Fig. 2. (left) Traffic light prior uncertainty ellipsoids increasing with transform chain length (at 99.99 % confidence). (right) Traffic Light 1 prior location and uncertainty from (left) projected into vehicle camera. Note there is an offset between predicted location and true location in the image. Turquoise - traffic light prior. Red - prior map projected into image. Green - prior uncertainty ellipsoids projected into image for 80%, 95%, 99.99% confidence.

A. Uncertainty Propagation

Without the use of expensive localisation sensors, there can be a significant level of uncertainty in the calculated traffic light positions, as shown in Fig. 2. Due to the uncertainty in the map construction process, traffic lights further from the vehicle typically exhibit higher positional uncertainties, reflecting the nature of uncertainty accumulation with increased distance. By explicitly quantifying the contributing uncertainties we can define high confidence real-space 3D search regions based around predicted traffic light locations in which to detect each traffic light.

Real-space uncertainty bounds for traffic light priors in detection range were determined according to the transform chain shown in Fig. 3 for 99.99 % confidence, resulting in a high confidence search region. To investigate the effect of low accuracy and low cost localisation systems, the uncertainty of the localisation transform between V and M_C can be increased to simulate the use of a lower-cost localisation system. At one extreme, the system provides an accurate estimate of the traffic light position within the image with uncertainty only due to the map construction process, significantly reducing the search region in the image; at the other extreme the search region grows to encompass the entire image, representing a naïve whole-image approach with an uninformative location prior.

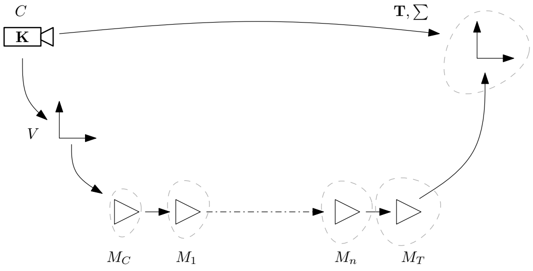


Fig. 3. Summary of composing traffic light prior projection, $\mathbf{T}$, and uncertainty, $\sum$, relative to camera C with camera matrix $\mathbf{K}$. First the transform is built relative to the vehicle frame V. Next, using a localisation system, the vehicle is localised relative to the current position in the map M_C, The prior map, M, is used to build a transform chain to the traffic light prior position, M_T, before applying the transform to the tagged traffic light prior, T. Uncertainty is composed at each step, resulting in a single prior projection transform, $\mathbf{T}$, and uncertainty, $\sum$.

B. Scale-Space Search Region

HOG feature descriptors are used to characterise and detect traffic lights as shown in Fig. 5. A 4x10 cell geometry was chosen to fit with UK traffic light regulations [15], allowing a tight crop around detected traffic lights, and to preserve the spatial characteristics of three vertical circle bulbs.

The high confidence 3D region was converted into camera image coordinates to enable a visual search. Camera projection of the uncertainty ellipsoid defines the maximum and minimum search locations (u, v). UK traffic light standards with ellipsoid extremity locations results in maximum and minimum estimated traffic light pixel sizes (s). Scales were generated at integer values in between, limiting the number of scales to 20 to reduce processing time. The resulting search region defines individual image patches (across the image search region at each scale) to visually compare to a

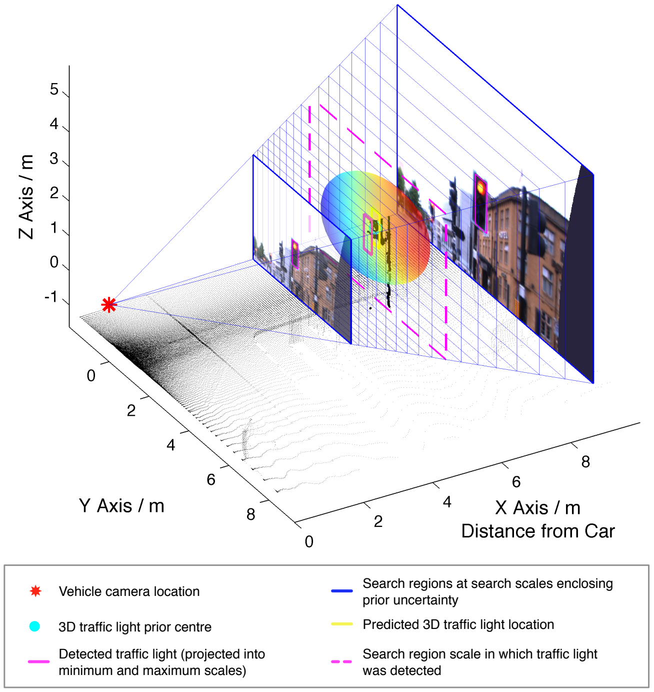


Fig. 4. Visualisation of the scale-space search region for Traffic Light 1 in Fig. 2 and the detected traffic light location. Black points represent the 3D prior map projected around the vehicle's predicted location. The search region is defined as to fully encapsulate the prior uncertainty for a traffic light at location (X, Y, Z) with a confidence of 99.99 %.

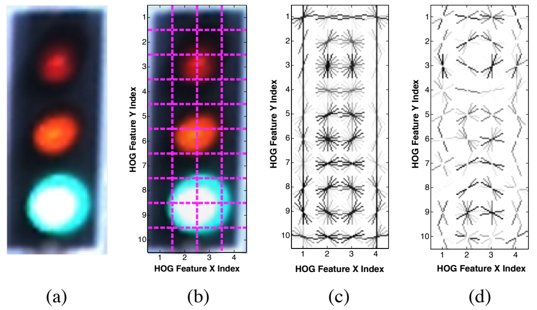


Fig. 5. Visualisation of the traffic light feature descriptor. A test image patch (a) is split into a cell geometry of 4x10 (b). HOG features are calculated for 9 unsigned orientation bins (c) before using a trained SVM model (d) to evaluate the likelihood of a traffic light.

traffic light template. When projected back, the search region fully encapsulates the 3D prior ellipsoid as shown in Fig. 4.

A linear SVM classifier is used to calculate the traffic light posterior $p(c|x)$. For each search scale, the SVM weights are convolved across a HOG representation of the search region, producing SVM scores at each test location.

Finally each search location is back-projected into 3D space and X, Y, Z components used to sample the prior uncertainty covariance $\sum$. The likelihood of a match at a given location is determined by the standard multivariate Gaussian PDF. Each final detection score is weighted by the corresponding occurrence prior, of which the highest score is deduced to be the detected traffic light location. This results in detector scores that incorporate both the classifier confidence (using HOG and SVM) and the occurrence prior (using the 3D location and associated map and localisation uncertainty).

C. State Interpretation

For traffic light state interpretation, an additional classification stage was added. A detected traffic light was split into three sections vertically and was classified by the image colour distribution in the Hue, Saturation, Lightness (HSL) colour space, which provides some robustness to illumination variance. This combination is beneficial as traffic light colours can vary between scenes for reasons such as fog, rain, shade and strong reflections, as shown in Fig. 6. States were classified as either Red, Amber, Green or Red & Amber.

V. Experimental Setup

The experimental data covered over 30 km of public roads with 102 unique traffic lights broken down into four datasets, one for training and three for evaluation, detailed in Table I. Two separate routes were used for data collection. The first covering 6.7 km of North Oxford with 44 traffic lights was recorded at three different times of day (shown in blue on Fig. 6), the second covering 10 km of Cowley with 58 traffic lights (shown in red). Night time use has not been explored however the underlying principle of confining and weighting the active search region is equally applicable.

TABLE I
Summary of Training and Testing Data

	Dataset	Frames Tagged	Traffic Lights on Route	Frames Traffic Light State / %			
				Red	Amber	Green	Red & Amber
Training	Dataset 1 - North Oxford Early Morning	1900	44	649	41	1008	202
Testing	Dataset 2 - North Oxford Early Morning	3071	44	1727	0	1099	245
	Dataset 3 - North Oxford Late Afternoon	4637	44	3106	42	1120	369
	Dataset 4 - Cowley Early Morning	1593	58	374	19	1116	84

The experimental platform was an autonomous Bowler Wildcat, equipped with a Point Grey Bumblebee2 stereo camera and Sick LMS151 laser scanner which were used to generate the 3D semantic prior map. Traffic light detection was carried out in a Point Grey Ladybug2 360° video camera.

As discussed in Sec. IV-A, additional localisation uncertainty was simulated for the sensors and methods listed in Tab. II in addition to the map uncertainty. Localisation systems simulated ranged from low-cost £35 consumer GPS devices [16] to tightly-integrated inertial systems costing upwards of £100,000 [6]. Uncertainties were derived from published statistics in manufacturer datasheets. The ‘Baseline’ method applies a trained detector over the whole scale space search region whereas ‘Whole Image with Prior’ improves on the ‘Baseline’ by utilising only the number of predicted traffic lights from the prior.

VI. Results

For evaluation we compare detected traffic lights locations and states in each image frame with a manually labelled ground truth. In Figs. 8(a) to 8(c), the left column shows traffic light detection precision at 99 % recall excluding state interpretation, and the right column displays the detection precision including state interpretation, referred to as pipeline precision. Precision for raw detection scores (without occurrence prior weighting) are shown for comparison and labelled ‘Raw Detection Scores’. ‘VO - Dense’ results show the effect of utilising a Dense HOG search over typical HOG implementations.

A. Detection and State Interpretation

Across the results three general relationships hold:

1) Traffic light detection precision is positively correlated to prior confidence.
2) Occurrence prior weighted detection scores always equal or outperform raw detection scores but have best effect with uncertain localisation.
3) Localisation with a variance up to 0.6 m, labelled SBAS (Satellite Based Augmentation System), produces detection precision comparable to highly accurate stereo VO results, indicating that higher accuracy (and therefore higher cost) solutions do not appreciably increase the detector performance.

Fig. 8(a) presents results for Dataset 2, the same route and time of day as training. As expected, detection performance is very high under similar conditions to training and displays

TABLE II
Localisation Methods Investigated

Method Label	x, y, z Additional Uncertainty Variance / m	Simulated Prior Source
Stereo VO	[0, 0, 0]	No additional uncertainty
RTK	[0.035, 0.035, 0.05]	Applanix POS-LV [6]
DGPS	[0.4, 0.4, 0.4]	Novatel SPAN-CPT [17]
SBAS	[0.6, 0.6, 0.6]	
GPS Raw	[1.2, 1.2, 1.2]	
GPS Consumer	[3.54, 3.54, 3.54]	SIRF Star III [16]
Whole Image with Prior	-	Number of traffic lights in image
Baseline	-	No Prior Information

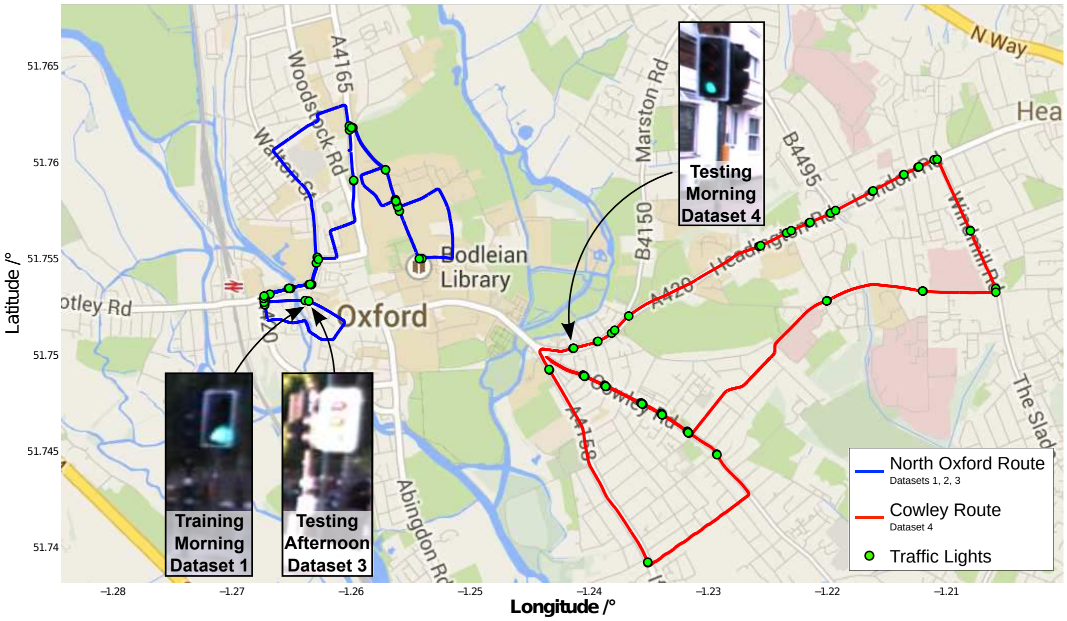


Fig. 6. Visual summary of 3D semantic prior map including traffic light locations, test routes used and example traffic light images. Traffic light appearance varied significantly during the course of the day, as shown by the example inset images for Dataset 1 and Dataset 3.

the three relationships listed. Occurrence prior weighting improved pipeline precision by up to 7 %.

Fig. 8(b) presents results for Dataset 3, the same route as training but different time of day. Detection precision is comparable to Dataset 2, however pipeline precision drops around 10 % due to the significant visual difference in the afternoon, shown in Fig. 6. Occurrence prior weighting improved pipeline precision by up to 15 %.

Fig. 8(c) presents results for Dataset 4, a different route to training, showing high performance on previously unseen traffic lights. Occurrence prior weighting improved pipeline precision by up to 15 %.

B. *Computation Time*

Fig. 7 presents average proportional traffic light detection times for each localisation method investigated, with each time normalised against the slowest method for that route. It is clear that as the confidence in the 3D traffic light prior increases, from the right to left, the detection time is reduced by up to a factor of 10 as the scale space search volume is smaller, motivating the use of higher quality priors. However, as with detection precision, the improvement of increased localisation accuracy after SBAS localisation is relatively small. The additional complexity of dense HOG (labelled VO - Dense) increases detection time significantly for minimal detection precision improvement over sparse HOG as shown in Fig. 8. With unoptimised MATLAB code, average Baseline performance of 5.23 s per frame on a 2.8 GHz Intel i7 is reduced by a factor of approximately 10 to 0.53 s with a strong VO or RTK prior.

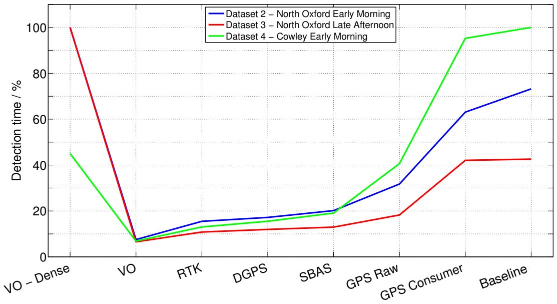


Fig. 7. Relative average traffic light detection time against localisation uncertainty. As 3D prior confidence increases, from right to left (excluding VO - Dense), the detection time decreases.

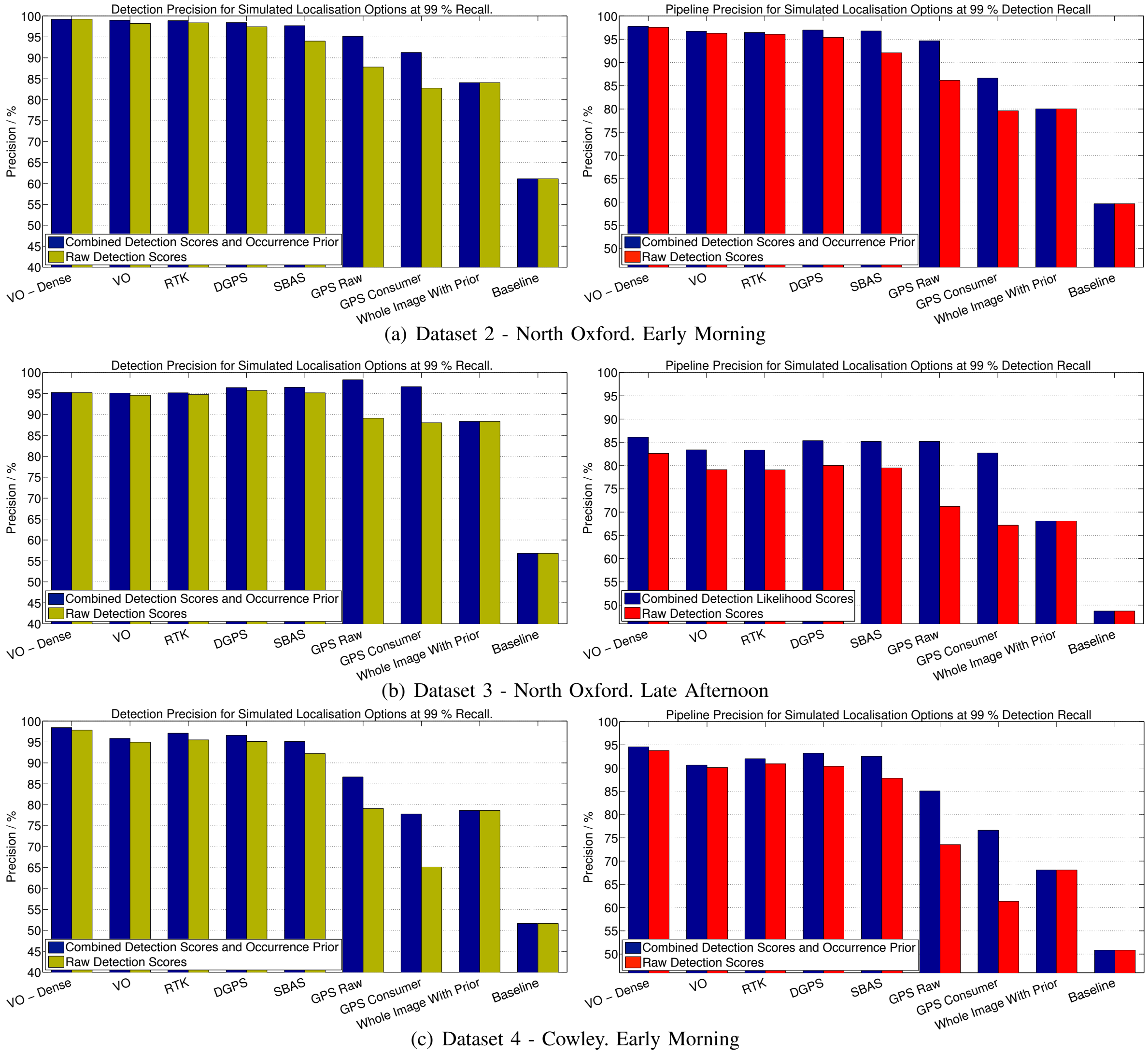
VII. Conclusions

This paper has presented a general framework to improve object detection performance given a 3D occurrence prior. Although a traffic light detection system has been implemented, the use of a widely applicable feature descriptor and flexible framework allows a variety of detection targets given a 3D semantic scene prior. No restrictions are imposed on collected data, other than quantifiable uncertainty, showing improved detection may be implemented on sensors already used for other purposes.

Crucially, our work demonstrates that weighting detection scores with occurrence priors improves detection performance under all tested conditions, ranging from large precision improvements of over 15 % on uncertain 3D priors to small improvements with accurate priors. Even simple prior knowledge in the form of the number of expected traffic lights per image increased pipeline precision by more than 20 %. This enables the use of lower-cost localisation for similar detection performance. Results have shown that for the purpose of traffic light detection given a 3D prior, there are minimal performance benefits with localisation uncertainty less than 0.6 m in our setup. By explicitly incorporating map and localisation uncertainty into the traffic light detection framework, we hope to provide reliable, accurate and lower-cost object detection for the autonomous vehicles and driver assistance systems of the future.

Acknowledgments

The authors would like to gratefully acknowledge support of this work by the European Community's Seventh Framework Programme under grant agreement no FP7-610603 (EUROPA2).



(a) Dataset 2 - North Oxford. Early Morning

(b) Dataset 3 - North Oxford. Late Afternoon

(c) Dataset 4 - Cowley. Early Morning

Fig. 8. Results shown for three test routes after morning training, comparing precision at 99 % recall with and without traffic light 3D occurrence prior weighting. (left) Traffic light detection. (right) Combined detection and state interpretation. Precision generally increases with localisation accuracy. Crucially, 3D occurrence prior weighting improves precision in all cases but proves more beneficial as priors become more uncertain (e.g. GPS Consumer), showing that high precision detection is feasible even with low-cost localisation hardware.

REFERENCES

[1] N. Fairfield and C. Urmson, "Traffic light mapping and detection," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5421–5426.

[2] J. Levinson, J. Askeland, J. Dolson, and S. Thrun, "Traffic light mapping, localization, and state detection for autonomous vehicles," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5784–5791.

[3] U. Franke, D. Pfeiffer, C. Rabe, C. Knoeppel, M. Enzweiler, F. Stein, and R. G. Herrtwich, "Making bertha see," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*. IEEE, 2013, pp. 214–221.

[4] F. Lindner, U. Kressel, and S. Kaelberer, "Robust recognition of traffic signals," in *Intelligent Vehicles Symposium, 2004 IEEE*. IEEE, 2004, pp. 49–53.

[5] H. Stubing, M. Bechler, D. Heussner, T. May, I. Radusch, H. Rechner, and P. Vogel, "sim td: a car-to-x system architecture for field operational tests [topics in automotive networking]," *Communications Magazine, IEEE*, vol. 48, no. 5, pp. 148–154, 2010.

[6] "Applanix POS LV,Position and Orientation System for Land Vehicles. http://www.applanix.com/media/downloads/products/specs/poslv_specifications12032012.pdf."

[7] J. Levinson, "Towards fully autonomous driving: Systems and algorithms," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE, 2011, pp. 163–168.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[10] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west." in *BMVC*, vol. 2, no. 3. Citeseer, 2010, p. 7.

[11] I. Cabani, G. Toulminet, and A. Bensrhair, "Color-based detection of vehicle lights," in *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*. IEEE, 2005, pp. 278–283.

[12] W. Churchill and P. Newman, "Experience-based Navigation for Long-term Localisation," *The International Journal of Robotics Research (IJRR)*, 2013.

[13] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[14] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[15] S. I. . N. 3113, "The traffic signs regulations and general directions 2002." Tech. Rep., 2002.

[16] "SiRF Star III, GPS Module. http://nz.apexelex.com/specs/modules_gps/MG-S02_v1.04.pdf."

[17] "NOVATEL SPAN-CPT, global navigation satellite system. http://www.novatel.com/assets/Documents/Papers/SPAN-CPT.pdf."