Describing Composite Urban Workspaces

Ingmar Posner, Derik Schroeter and Paul Newman

Abstract— **In this paper we present an appearance-based method for augmenting maps of outdoor urban environments with higher-order, semantic labels. Our motivation is to increase the value and utility of the typically low-level representations built by contemporary SLAM algorithms. A supervised learning scheme is employed to train a set of classifiers to respond to common scene attributes given a mixture of geometric and visual scene information. The union of classifier responses yields a composite description of the local workspace. We apply our method to three large data sets.**

I. INTRODUCTION

Localisation and Mapping frameworks have reached a level of maturity such that a vehicle can traverse and map substantial workspaces. The run-time complexity of state estimation algorithms is no longer the primary bottleneck. However, the maps produced are typically agglomerations of laser points or an arrangement of geometric primitives (often simply points, lines and planes). Such representations only have a limited discriminative capacity and fail to adequately represent the subtleties of complex environments. As a consequence, data association, pivotal to the construction of consistent maps, remains an open problem — perhaps the Achilles' heel of the research domain.

Appearance-based techniques developed in the computer vision domain have emerged as a valuable complement to standard SLAM solutions [1], [2]. An example is the robust closing of large loops in a vehicle's trajectory using an appearance-based visual loop-closing engine [3]. The salient point here is that the data-association problem can be addressed without metric reasoning — considering what things look like as opposed to where they appear to be. The annotation of common SLAM maps by semantic information seems a natural extension of this notion.

Our goal, therefore, is to add value to maps built by SLAM algorithms by augmenting them with higher-order, semantic labels. Such labels are vastly more descriptive than the geometric primitives used previously and thus contribute considerably to a correct data association. In this paper we achieve this by using both scene appearance and geometry to produce a composite description of the local area in urban settings. Outdoors, we use a 3D laser scanner to sense the local workspace geometry and a camera to capture its visual appearance. In combination these two sensors provide a rich source of information with which to characterise different aspects of the local area. In particular, we will focus on describing regions of the ground plane, the surface type of any walls in view and the presence or otherwise of cars and foliage. The geometric and visual properties of a particular scene are passed through a bank of classifiers each trained to respond to a given scene attribute — like pavement, tarmac or bush. The combination of all positive classifications yields a composite description of the scene in question, for example, "Path and Grass and Foliage" or "Road and Brick-Wall and Car" (Fig. 1). The classifiers process a mixture of geometric and appearance information which is extracted in the following way. Firstly, using the 3D laser data, planar patches are extracted and the normals recorded. Then, each constituent laser point in a patch is back projected into the camera image and neighbourhood parametrised (via a colour histogram) and recorded along with its image coordinates. Training is done using hand-labelled data.

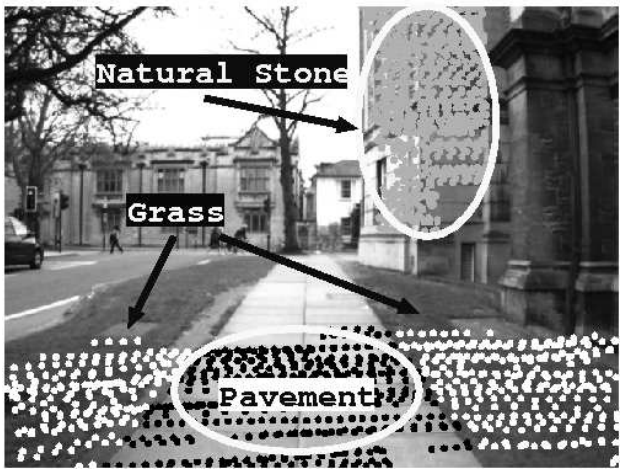

Fig. 1. Labels for a typical urban scene

The next section gives a brief overview of related works. Section III describes the data used. A motivation of our choice of workspace labels is given in Section IV. This is followed by a detailed description of the features used and the data processing applied in Section V. The learning of appropriate classifiers is outlined in Section VI. The applicability of the presented approach to urban settings is demonstrated in Section VII. We conclude with a brief summary and discussion of future work in Section VIII.

II. RELATED WORK

The extraction of semantic information from sensor data has received much attention in recent years and the amount of relevant literature is substantial. In the computer vision domain, approaches to appearance-based scene and object

The work reported in this paper was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence.

The authors are with the Robotics Research Group, Dept. Engineering Science, Oxford University, Oxford, OX1 3PJ, United Kingdom {hip, ds, pnemwan}@robots.ox.ac.uk

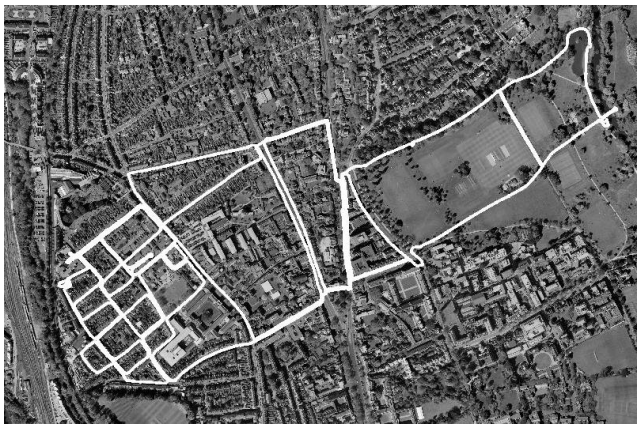

Fig. 2. An aerial map of the *Jericho* data set (13.2 km, 16000 images). The vehicle's trajectory is marked in white.

classification include unsupervised statistical methods applied to bags of features, both including [4] and excluding [5] position within an image. In the robotics domain, recent developments include the classification of traversable regions from both laser and image data [6], the unsupervised partitioning of outdoor workspaces using image similarity [7] and the classification of 2D laser data into types of indoor scenes using boosting [8]. Contextual information was used explicitly in [9] by way of a model based on relational Markov networks to learn classifiers from segment-based representations of indoor workspaces. In [10] 3D laser data is segmented to detect cars and classify terrain using Graph Cut applied to a Markov Random Field (MRF) formulation of the problem. The performance of the MRF framework is compared to that obtained using (voted) support vector machine classification. In a sense this work is most closely related to our approach in that we also employ support vector machines to classify laser data. However, in combining information from two complimentary sensors – geometry and appearance – our approach gains the capacity of providing *more detailed* workspace descriptions such as the surface-type of building(s) encountered or the nature of ground traversed.

III. URBAN DATA

The work presented in this paper makes use of three extensive data sets spanning nearly 18 km of track gathered with an ATRV mobile platform. The robot is equipped with a colour camera mounted on a pan-tilt unit, an inertial sensor (XSens) as well as a GPS sensor and odometry from wheel encoders. The camera records images to the left, the right and the front of the robot in a pre-defined pan-cycle triggered by vehicle odometry at 1.5 m intervals. 3D laser data are acquired using a standard 2D SICK laser range finder (75 Hz, 180 range measurements per scan) mounted in a reciprocating cradle driven by a constant velocity motor. Data recorded from all sensors are time-stamped on arrival.

Data were gathered in three different locations: *Jericho/Oxford* (13.2 km, 16,000 images, Fig. 2), *Edinburgh* (1.3 km, 3561 images) and the *Oxford Science Park* (3.3 km, 8536 images, Fig. 3).

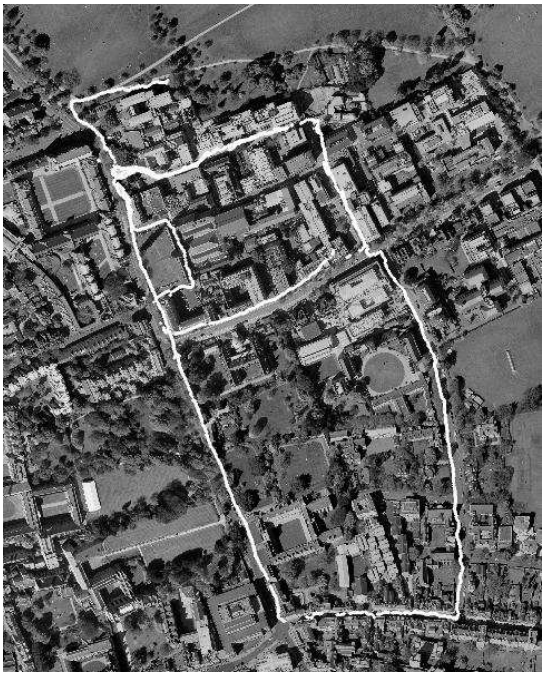

Fig. 3. An aerial map of the *Oxford Science Park* data set (3.3 km, 8536 images). The vehicle's trajectory is marked in white.

IV. WORKSPACE CLASSES IN URBAN ENVIRONMENTS

When navigating in an urban context a higher-order knowledge of the environment is indispensable: self-preservation dictates avoidance of highly dynamic regions such as roads; robust localisation depends on distinguishing features beyond the recognition of ubiquitous general objects such as 'ground', 'wall' or 'house'. This necessity motivates the definition of classes and the closely linked selection of features in this work. Intuitively, in an urban environment places can be distinguished by the type of ground that is present, the colour and texture of surrounding houses (or, more appropriately, of surrounding walls) and the presence or absence of other features such as bushes or trees. The detection of cars (moving or stationary) is also beneficial. These considerations give rise to the classes defined in Table I.

TABLE I

WORKSPACE CLASSES.

Class Name	Description
Wall Structure	
Brick	red or yellow brick
Nat. Stone	natural stone, sandstone
Concrete	modern (e.g. concrete, glass)
Plastered	plastered, painted
Ground	
Pavement	tiled, patched
Path	sand / dirt / gravel
Grass	grass
Tarmac	common road, pavement
Nature	
Bush or Foliage	bushes and parts of trees
Miscellaneous	
Vehicle	cars or vans

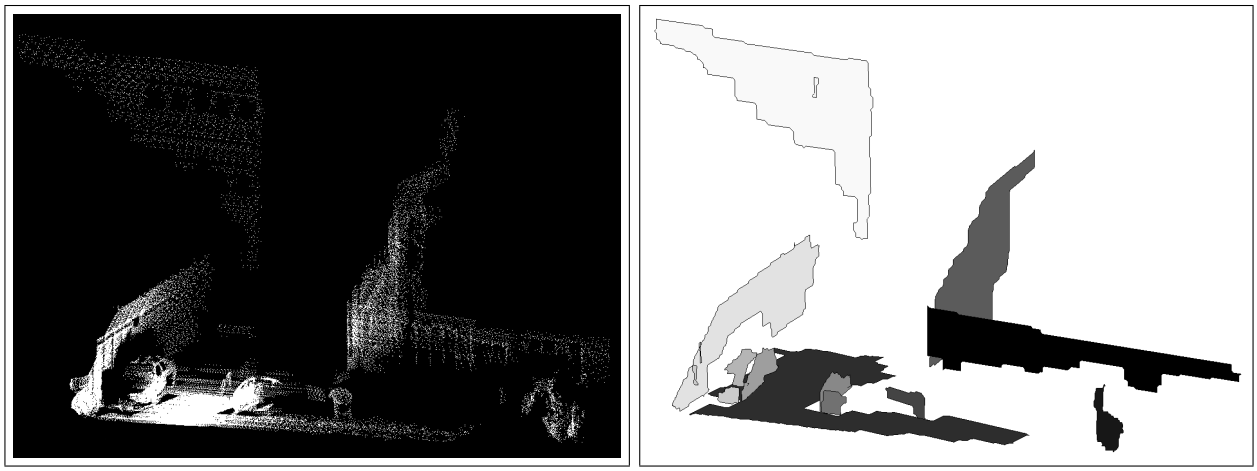

Fig. 4. The left image shows an original 3D laser scan, the right depicts its approxmation by planar patches as generated by the segmentation algorithm.

V. FEATURE EXTRACTION

The classes defined in Table I suggest both visual (colour and texture) and 3D geometrical features. Our vehicle is equipped with a 3D laser scanner, which supplies direct measurements of geometry. Knowledge of the intrinsic as well as the extrinsic (wrt the laser range finder) camera parameters allows a meaningful combination of laser measurements and image data: each laser measurement can be augmented with local colour and texture information. Starting with a colour image and a 'cloud' of laser measurements, an appropriate feature vector can be compiled incorporating both 3D geometrical (laser) and appearance (camera) features. The choice of features from the two modalities and their extraction is described in the following.

Laser Features. Using the time at which the colour image was taken as reference, 3D laser points are accumulated over a time window of length Δt into the past. Thus, a 3D point cloud is assembled which represents the original scene subject to the colour image. The structural and ground classes in Table I can be approximated geometrically with a planar model. Therefore, the 3D laser data associated with an image were segmented into planes following a divide-and-conquer approach outlined in [11]: a given point cloud is discretised into cubic cells and planes are fitted locally using RANSAC [12]. Plane segments for which the support (i.e. the number of inliers) is less than a threshold, are discarded. Amongst the survivors, planes obtained in neighbouring cells are merged according to two constraints relating to relative surface orientation and translation. The merging criteria for orientation and translation are specified as:

$$| \mathbf{n_i} \cdot \mathbf{n_j} | > \arccos(\alpha_{max}) \quad and \quad \frac{1}{2}(d_{ij} + d_{ji}) < d_{max}$$

$\mathbf{n_i}$ and $\mathbf{n_j}$ denote the plane normals in cells *i* and *j* and '$\cdot$' denotes the scalar product. d_{ij} and d_{ji} denote the distances from the centre of gravity of one plane to its orthogonal projection onto the other plane (Fig. 5). α_{max} and d_{max} denote an angle threshold and a distance threshold, respectively. Finally, merged plane patches are kept if they comprise more than N_{min} laser points. A typical result of this segmentation process is shown in Fig. 4.

Currently only the absolute cosine distance between a plane normal and the normal of the ground plane is used as a 3D geometric feature.

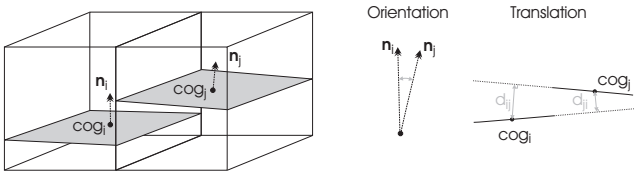

Fig. 5. The plane-merging constraints for orientation and translation for two adjacent cells *i* and *j*. $\mathbf{n}$ and *cog* denote the plane normals and the centres of gravity, respectively.

Appearance Features. The processing pipeline as described above provides 3D laser points which lie on planes fitted to the original laser data, covering the scene depicted in the image *and beyond*. Visual features can only be extracted for laser points which fall within the field of view of the camera. Thus, irrelevant laser data are filtered out using a standard frustum culling technique. The remaining laser points are projected into the image (Fig. 6). Using these projections as 'points of interest', appearance features are calculated over a fixed-size (15×15 pixels) local neighbourhood in the image. Colour and texture were deemed the most important visual features as they provide information about the material a surface is made of. Colour is represented by local hue and saturation histograms (15 bins). A very basic additional texture feature was computed for each colour channel simply by taking its variance. The use of more advanced texture descriptors derived from Gabor filters, for example, was considered but decided against at this stage in favour of simplicity.

In addition to these visual features the normalised 2D position of the projections, as proposed by Hoiem et al

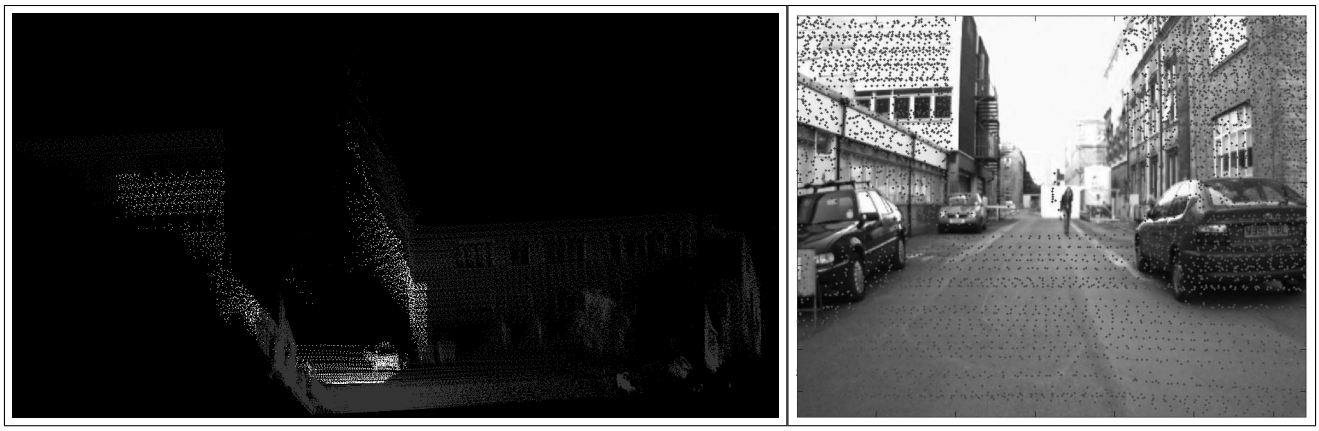

Fig. 6. Camera-laser cross-calibration: a typical 3D laser point-cloud (left). Laser points within the camera frustum are highlighted (white) and projected into the corresponding camera image (right).

[13], was also added to the feature vector. The motivation is that, since the camera only rotates around the vertical axis, observations of the ground plane are more likely to appear in the lower part of an image whereas walls of buildings extend into the upper part.

A flowchart of this processing pipeline for feature extraction is given in Fig. 7. It currently runs offline as a Matlab implementation at about four seconds per image. The features extracted are summarised in Table II. It remains the task of assigning a certain semantic label to each of the laser points based on this information. This is a classical machine learning problem and will be addressed in Section VI.

1) For image I taken at pose x_I and time t_I:
   (a) Obtain 3D laser data (L, t_L) temporally close to t_I, i.e. $t_I - \Delta t < t_L < t_I$
2) Segment planar patches from 3D point cloud, keep patches that comprise more than N_{min} points.
   Note: N_{min} is different from the inlier threshold used for RANSAC.
3) Filter out 3D points that do not lie within the viewing frustum of the camera (frustum culling).
4) For each of the remaining 3D points:
   (a) Assign the 3D geometric features from the respective plane patch (Table II).
   (b) Project the 3D point into the image.
   (c) Compute 2D geometric, colour and texture features (Table II) from a local neighbourhood.

Fig. 7. The processing pipeline employed for feature extraction.

TABLE II

GEOMETRIC AND APPEARANCE-BASED FEATURES USED FOR CLASSIFICATION

Feature Descriptions	Dimensions
3D Geometry Orientation of surface normal of local plane	1
2D Geometry Location in image: mean of normalised x and y	2
Colour HSV: hue & sat. histograms (15 bins)	30
Texture HSV: hue & sat. variance in local neighbourhood	2

VI. CLASSIFICATION

For classification we chose a chain of support-vector machines (SVMs) with a Gaussian kernel[1]. SVMs are based on a linear discriminant framework which aims to maximise the margin between two classes. They are a popular choice since the model parameters are found by solving a convex optimisation problem. This is a desirable property since it implies that the final classifier is guaranteed to be the best feasible discriminant given the training data. SVMs are inherently binary classifiers. In this work, multi-class classification is performed by training a chain of binary classifiers – one for each class – as one-versus-all [15].

TABLE III

CLASSIFIER PERFORMANCE STATISTICS ON A TEST SET [%] .

Classifier	Accuracy	Precision	Recall
Grass	98.5	99.4	97.5
Paved	86.7	89.0	83.7
Dirt	86.6	93.7	78.5
Tarmac	93.8	94.8	92.5
Brick Wall	89.9	94.7	84.5
Nat. Stone Wall	90.6	94.0	86.9
Concrete Wall	83.7	90.0	75.8
Plastered Wall	85.1	80.5	92.8
Bushes/Foliage	95.2	97.8	92.5
Vehicles	91.3	96.2	85.9

Training. SVM training was conducted using the *Jericho* data set. The appropriate kernel width and the regularisation parameter (i.e. the tolerance for misclassifications) were determined using a grid-search over a section of the parameter space. The grid-search was conducted with 6,000 training points and 4,000 test points per class. The data were balanced so that training was conducted at an equal ratio of positive to negative examples. The parameter-set resulting in the highest overall classification accuracy was chosen for each class (see Table III) and the corresponding classifier was re-trained using the entire training set of 10,000 data points. A good indication of the generalisation performance of these classifiers across data gathered in independent locations and

[1]SVM training and classification were performed using SVMLight [14].

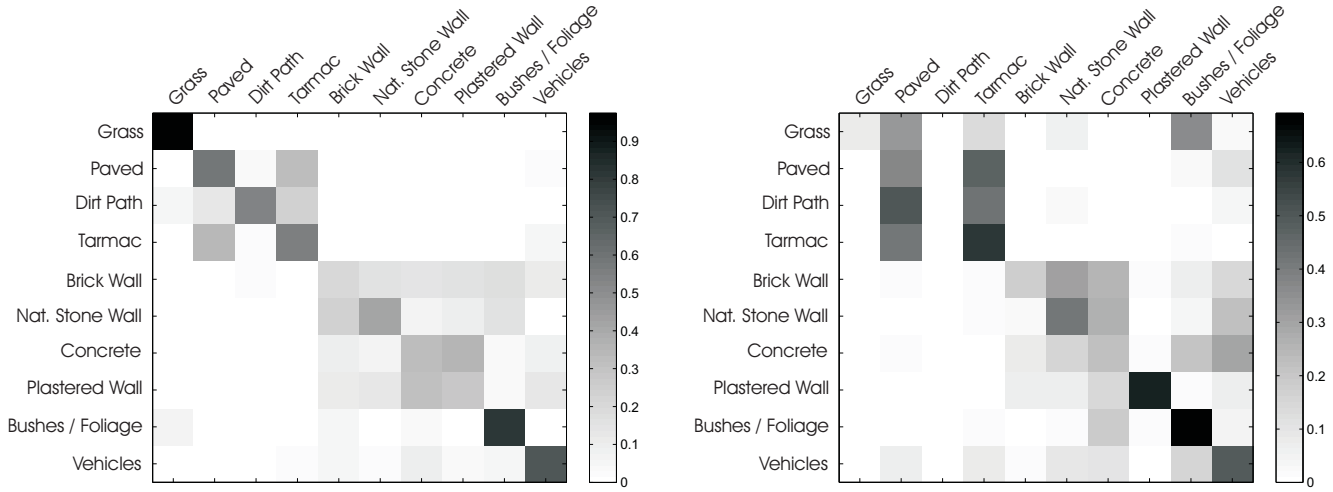

Fig. 8. A graphical representation of the normalised confusion matrices for the Oxford Science Park data set (left) and the Edinburgh data set (right).

under vastly different conditions can be gained by inspection of the confusion matrices in Section VII.

Classification. The predicted class of a datum is that for which it is classified with the greatest margin [15]. If none of the classifiers in the chain associate the datum with their respective class, the data point remains unclassified.

VII. RESULTS

The previous section outlined the training of a chain of binary classifiers using the *Jericho* data set. The generalisation performance of these classifiers was tested using labelled data from both the *Oxford Science Park* and the *Edinburgh* data sets (ca. 52,300 and 38,700 data, respectively). A graphical representation of the confusion matrices for both data sets is given in Fig. 8. Full details are given in Tables IV and V.

The matrix originating from the *Oxford Science Park* data is dominated by high values on the diagonal. Grass, bushes/foliage and vehicles are classified with consistently high precision. Striking is the consistent block-separation between ground and non-ground (walls, bushes/foliage and vehicles). This is attributed to the features describing the orientation of plane patches and the location of laser points within an image. Types of terrain other than grass are harder to distinguish between. Paved or patched walkways, dirt paths and roads/pavements with a tarmac surface can be similar in colour and texture, giving rise to confusion. Nevertheless, the majority of classifications are consistently correct. Greater confusion can be observed amongst the different types of walls, where a similar argument applies with regards to colour. Block-cohesion can be observed amongst brick and natural-stone walls as well as concrete and plastered walls. This may be attributed to a difference in texture.

The matrix originating from the *Edinburgh* data exhibits a broadly similar structure but is considerably more noisy. This is attributed to the sub-optimal lighting conditions prevailing while the data was gathered, since it may have given rise to higher variability in feature values describing colour and texture.

The consistency of the classification results can be further emphasised by combining conceptually related classes for which the current combination of descriptive features does not allow for robust classification. For example, the *Oxford Science Park* data (Fig. 8) suggest a block-cohesion between the 'Concrete' and the 'Plastered Wall' classes as well as the 'Brick Wall ' and the 'Nat. Stone Wall' classes. This is most likely due to texture (and possibly colour) similarities within those groups, yet not across. Fig. 10 depicts the confusion matrices for the respective data sets with two meta classes 'Textured Wall' and 'Plain Wall'. These represent the combined class pairs 'Brick Wall ' and 'Nat. Stone Wall', and 'Textured Wall' and 'Plain Wall', respectively. The dominance in the diagonal has increased.

So far, discrete laser points sampled from a continuous world have been classified independently, thus discarding all information about the spatial cohesion of structures and objects. Taking this information into account leads to an intuitive extension: the smoothing of individual classification results by majority vote of laser points constituent to the same plane patch. As a preliminary investigation, this technique was applied to the data obtained from the *Oxford Science Park*. The resulting confusion matrix (Fig. 9) exhibits a more pronounced diagonal and less noise. This suggests that such 'spatial' smoothing may indeed improve overall classification performance. A typical example of an actual classified scene where a majority vote scheme is applied is given in Fig. 1.

VIII. CONCLUSIONS AND FUTURE WORKS

In this paper we present an appearance-based method of augmenting maps of outdoor urban environments with local scene labels. The approach is based on a chain of binary classifiers labelling individual laser data according to their origin. Laser points are characterised by both 3D geometric

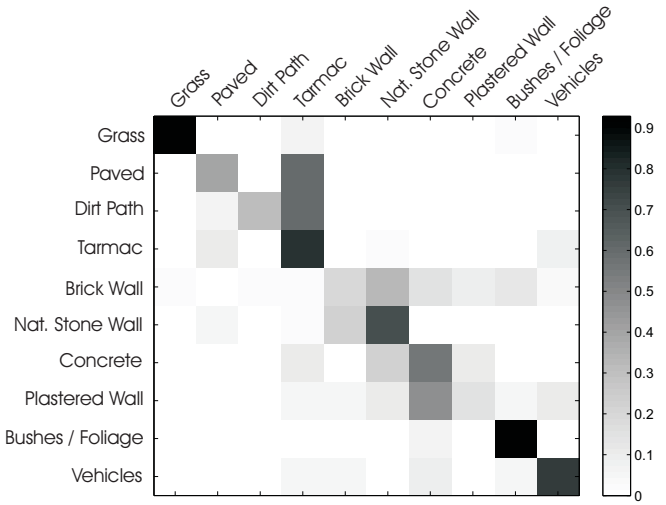

Fig. 9. The normalised confusion matrix for the 'spatially smoothed' data of the *Oxford Science Park* data set.

data and visual cues obtained from monocular vision. The generalisation performance of the classification scheme is sufficient to consistently separate different types of terrain and walls, including bushes and foliage. The system also has a capacity to recognise common objects such as cars and vans. The results suggest that this approach can be extended towards the smoothing of individual classification results by taking into account the spatial cohesion underlying the point cloud. However, such a scheme relies on an automatic separation of plane patches into surfaces of different types. Currently, this is beyond the segmentation-scheme applied here and is subject to further work.

In the future, attention will also focus on an evaluation of the feature set used. At this point, no comment can be made on the relative importance of individual features to the classification process. Though the classification performance is satisfactory, it may well transpire that our system would benefit from, for example, more advanced texture features or more elaborate geometric features.

Furthermore, the use of an inherently binary classification framework in a one-verses-all configuration comes with a caveat: the possibility of individual classifiers assigning an input to multiple classes simultaneously is addressed using a 'winner-takes-all' heuristic where the 'winner' is the classification resulting in the greatest margin. Even though satisfactory results are obtained in practise, there is no guarantee that the real valued quantities representing the margins for different classifiers will have appropriate scales [16]. In future, this will be addressed by investigating alternative classification frameworks such as relevance vector machines which do not suffer this limitation.

IX. ACKNOWLEDGEMENTS

The work reported in this paper was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence. The authors gratefully acknowledge the support of the members of the Robotics Research Group who have tirelessly labelled our laser data.

REFERENCES

[1] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," *Proceedings of International Conference on Robotics and Automation*, 2000.

[2] J. Porta and B. J. A. Kroese, "Appearance-based concurrent map building and localization," *Robotics and Autonomous Systems*, vol. 54, no. 2, pp. 159–164, 2005.

[3] P. Newman, D. Cole, and K. Ho, "Outdoor SLAM using visual appearance and laser ranging," *IEEE International Conference on Robotics and Automation*, May 2006.

[4] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories from Google's Image Search," in *Proc. of the Int. Conference on Computer Vision*, 2005.

[5] A. Bosch, A. Zisserman, and X. Munoz, "Scene Classification via pLSA," in *Proc. of the European Conference on Computer Vision*, 2006.

[6] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney, "Winning the darpa grand challenge," *Journal of field Robotics*, 2006.

[7] I. Posner, D. Schroeter, and P. M. Newman, "Using scene similarity for place labelling," in *Proc. of the Int. Symposium on Experimental Robotics*, 2006.

[8] C. Stachniss, O. Martínez-Mozos, A. Rottmann, and W. Burgard, "Semantic labeling of places," San Francisco, CA, USA, 2005.

[9] B. Limketkai, L. Liao, and D. Fox, "Relational object maps for mobile robots." in *IJCAI*, L. P. Kaelbling and A. Saffiotti, Eds. Professional Book Center, 2005, pp. 1471–1476.

[10] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Y. Ng, "Discriminative learning of Markov random fields for segmentation of 3D scan data." in *CVPR (2)*. IEEE Computer Society, 2005, pp. 169–176.

[11] J. Weingarten, G. Gruener, and R. Siegwart, "A fast and robust 3D feature extraction algorithm for structured environment reconstruction," in *Proc. of the 11th International Conference on Advanced Robotics (ICAR)*, 2003.

[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[13] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, June 2006.

[14] T. Joachims, "Making large-scale support vector machine learning practical," pp. 169–184, 1999.

[15] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, August 2006.

TABLE IV

CONFUSION MATRIX FOR THE OXFORD SCIENCE PARK DATA SET.

		Ground truth class labels									
		Grass	Paved	Dirt	Tarmac	Brick	Natural Stone	Concrete	Plastered	Bushes	Cars
Classification	Grass	4320	31	38	21	0	0	0	0	13	10
	Paved	59	3178	239	1808	2	37	3	33	5	108
	Dirt	398	917	4078	1771	18	68	88	27	17	89
	Tarmac	19	733	33	1175	2	23	12	4	2	128
	Brick	60	80	458	99	3072	2526	2215	2362	2601	1660
	Natural Stone	9	20	0	12	921	1682	301	414	611	22
	Concrete	9	15	2	14	341	231	1121	1221	106	300
	Plastered	0	1	0	6	340	342	854	792	110	357
	Bushes	111	8	8	7	76	7	60	3	1310	17
	Cars	2	15	12	69	158	58	321	120	181	2224
	Unclassified	2	2	131	6	65	21	4	6	22	72
	Ground Truth	4999	5000	4999	4999	5002	4997	4997	4998	4999	5000

TABLE V

CONFUSION MATRIX FOR THE EDINBURGH DATA SET.

		Ground truth class labels									
		Grass	Paved	Dirt	Tarmac	Brick	Natural Stone	Concrete	Plastered	Bushes	Cars
Classification	Grass	19	80	0	33	0	14	1	0	88	7
	Paved	0	660	0	831	4	13	9	0	40	202
	Dirt	6	3088	0	2584	3	133	28	0	33	208
	Tarmac	0	721	0	1004	0	0	0	0	19	7
	Brick	0	154	0	192	1723	3100	2504	166	698	1402
	Natural Stone	0	0	0	17	49	656	404	15	62	353
	Concrete	0	29	0	15	191	386	568	43	522	748
	Plastered	0	2	0	0	388	392	833	3414	89	402
	Bushes	0	19	0	41	11	38	380	30	1415	110
	Cars	0	97	0	108	20	120	133	3	217	663
	Unclassified	0	146	0	169	34	80	99	5	10	223
	Ground Truth	25	4998	0	4995	2427	5000	4999	3679	3235	4347

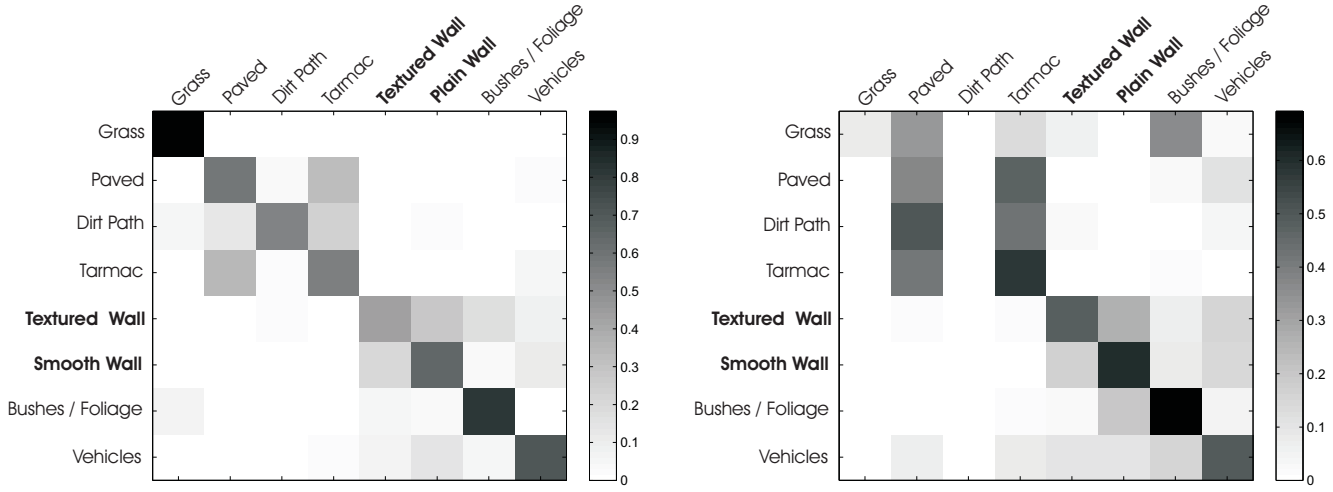

Fig. 10. A graphical representation of the normalised confusion matrices for the *meta classes* of the Oxford Science Park data set (left) and the Edinburgh data set (right).