2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

Automatic Dense Visual Semantic Mapping from Street-Level Imagery

Sunando Sengupta[1] Paul Sturgess[1] L'ubor Ladický[2] Philip H. S. Torr[1]
{ssengupta, paul.sturgess, philiptorr}@brookes.ac.uk, lubor@robots.ox.ac.uk
[1]Oxford Brookes University [2]University of Oxford

***Abstract*— This paper describes a method for producing a semantic map from multi-view street-level imagery. We define a semantic map as an overhead, or bird's eye view of a region with associated semantic object labels, such as car, road and pavement. We formulate the problem using two conditional random fields. The first is used to model the semantic image segmentation of the street view imagery treating each image independently. The outputs of this stage are then aggregated over many images to form the input for our semantic map that is a second random field defined over a ground plane. Each image is related by a simple, yet effective, geometrical function that back projects a region from the street view image into the overhead ground plane map. We introduce, and make publicly available, a new dataset created from real world data. Our qualitative evaluation is performed on this data consisting of a 14.8 km track, and we also quantify our results on a representative subset.**

I. Introduction

In this paper, we introduce a computer vision based system exploiting visual data for semantic map generation. Images can be acquired at a low cost and provide much more informative features than laser ranger scans. This makes our application unique compared to those that use other modalities for recognition and semantic mapping [4], [9], [17], [18], [19]. Another important motivation for using visual data is that it can be captured at high frequency and resolution, allowing us to classify all the individual pixels in the image rather than a sparse set of returns from a laser scan [20], [21]. We call this *dense visual semantic mapping*.

The ability to automatically create outdoor mapping data with semantic information is valuable for many robotic tasks such as determining drivable regions for autonomous vehicles [7], navigating structured environments [25], and interacting with objects [5].

Our method for performing semantic mapping combines two distinct conditional random fields (CRF). The first CRF works with a stereo pair (or more) of images taken from synchronized cameras on the vehicle, feeding into and updating the second that globally optimizes the semantic map. We chose to use CRFs as promising results have been demonstrated on street-level imagery [23], [15]. The two CRFs' are linked via a homography. The two stage process enables us to model spatial contextual relations in both the image domain as well as on the ground plane. CRFs are flexible probabilistic models, so that other suitable modalities of data could be included into our framework in the future.

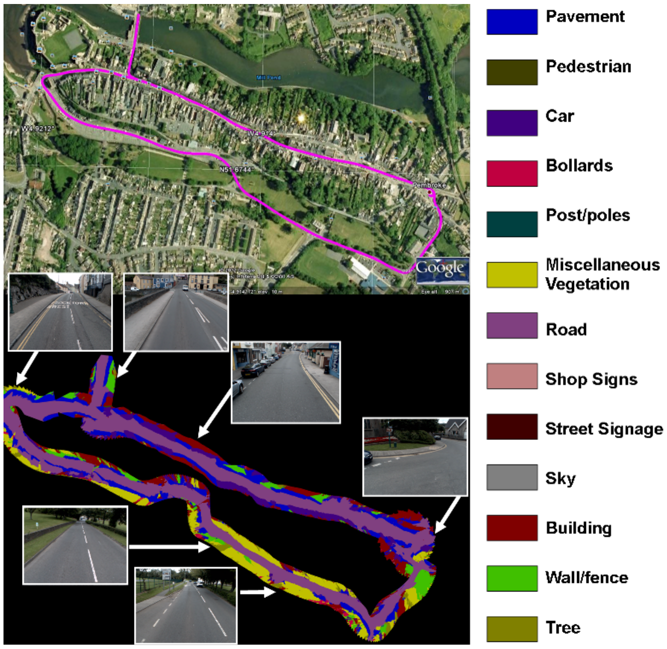

Fig. 1. *Dense Semantic Mapping.* Top: Visualization of the route taken by a specialized van to capture this street level imagery that we use to evaluate our method. The Semantic map (bottom) is shown for the corresponding track along with some example street view images. The palette of colours shows the corresponding class labels. Best viewed in colour.

We evaluate our method on a subset of data that covers all the roadways of the United Kingdom captured by YottaDCL [8]. Their capturing process is performed by a specialized vehicle, see Fig. 5, fitted with two frontal, two sideways and two rear cameras. The vehicle also has GPS and odometry devices so that its location and camera tracks can be determined accurately. Our qualitative evaluation is performed on an area that covers 14.8 km of varying roadways. We have hand labelled a relatively small, but representative set of these images with per-pixel labels for training and quantitative evaluation. All 6 views of our 14.8km track are publicly available along with the annotation data[1]. We hope that this stimulates further work in this area. This type of data along with our formulation allows us to aggregate semantic information over many frames, providing a robust classifier. Some results are shown in Fig. 1 along with the associated Google Earth [11] satellite view and our street-level images. For a graphical overview of the proposed system see Fig. 2.

In summary our main contributions are:

- We introduce the problem of dense semantic mapping

[1] Please see http://cms.brookes.ac.uk/research/visiongroup/projects.php

978-1-4673-1736-8/12/S31.00 ©2012 IEEE

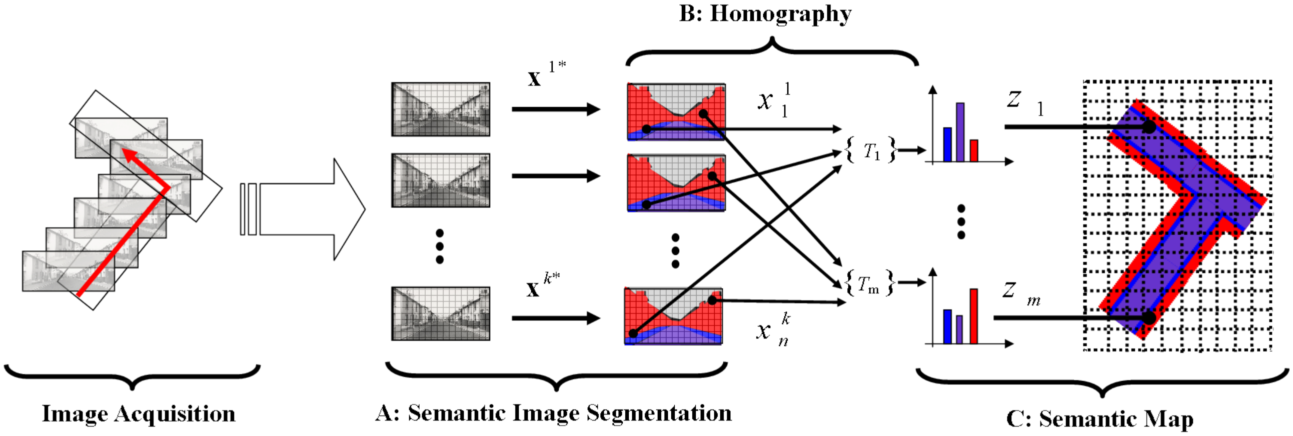

Fig. 2. *Overview of our Algorithm* (a) Each image is treated independently and each pixel in the image is classified, giving a labelling $\mathbf{x}^*$ (§II-A). (b) A homography gives the relationship between the pixel on the ground plane to a pixel in a sub-set of the images, where the ground pixel is visible (§II-B). T_m denotes the set of the image pixels that correspond to the ground plane pixel z_m. (c) The CRF producing a labelling of the ground plane, or a semantic map (§II-C).

from multiple view street level imagery: §I.

- We define a CRF model to address the problem: §II.
- We make publicly available a street-level dataset with multiple views, camera tracks and partially hand labelled ground truth to stimulate research in the area: §III.

II. Conditional Random Field Model

We model the problem of dense visual sematic mapping using two CRFs. In this section we fist introduce notations [23] and then define both models and a homography that links them together.

Conditional Random Field Notation: Consider a set of random variables $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, where each variable $X_i \in \mathbf{X}$ takes a value from a pre-defined label set $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$. A labelling $\mathbf{x}$ refers to any possible assignment of labels to the random variables and takes values from the set $\mathbf{L} = \mathcal{L}^N$. The random field is defined over a lattice $\mathcal{V} = \{1, 2, \dots, N\}$, where each lattice point, or pixel, $i \in \mathcal{V}$ is associated with its corresponding random variable X_i. Let $\mathcal{N}$ be the neighbourhood system of the random field defined by sets $\mathcal{N}_i, \forall i \in \mathcal{V}$, where $\mathcal{N}_i$ denotes the set of all neighbours (usually the 4 or 8 nearest pixels) of the variable X_i. A clique c is defined as a set of random variables $\mathbf{X}_c$ which are conditionally dependent on each other. The posterior distribution $\Pr(\mathbf{x}|\mathbf{D}) = \frac{1}{Z}\exp(-\sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c))$ over the labellings of the CRF is a *Gibbs* distribution, where the term $\psi_c(\mathbf{x}_c)$ is known as the potential function of the clique c, and $\mathcal{C}$ is the set of all the cliques. The corresponding Gibbs energy $E(\mathbf{x})$ is given by: $E(\mathbf{x}) = -\log \Pr(\mathbf{x}|\mathbf{D}) - \log Z$, or equivalently $E(\mathbf{x}) = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$ The most probable or maximum a posteriori labelling $\mathbf{x}^*$ of the CRF is defined as: $\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathbf{L}} \Pr(\mathbf{x}|\mathbf{D}) = \arg\min_{\mathbf{x} \in \mathbf{L}} E(\mathbf{x})$.

A. Semantic Image Segmentation

For this part of our semantic mapping system we use an existing computer vision system [14] that is competitive to state-of-the art for general semantic image segmentation [10] and leading the field in street scene segmentation within the computer vision community [23]. We give a brief overview of their method here with respect to our application. Our label set is $\mathcal{L}$ = {pavement, building, road, vehicle, tree, shop sign, street-bollard, misc-vegetation, pedestrian, wall-fence, sky, misc-pole, street-sign}. We used the associative hierarchical CRF [14] which combines features and classifiers at different levels of the hierarchy (pixels and superpixels). The Gibbs energy for a street-level image is

$$E^S(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i^S(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^S(x_i, x_j) + \sum_{c \in \mathcal{C}} \psi_c^S(\mathbf{x}_c). \tag{1}$$

Unary potential: The unary potential ψ_i^S describes the cost of a single pixel taking a particular label. It is learnt as the multiple-feature variant [14] of TextonBoost algorithm [22].

Pairwise potential: The 4-neighbourhood pairwise term ψ_{ij}^S takes the form of a contrast sensitive Potts model:

$$\psi_{ij}^S(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ g(i,j) & \text{otherwise,} \end{cases} \tag{2}$$

where the function $g(i,j)$ is an edge feature based on the difference in colours of neighbouring pixels [1], defined as:

$$g(i,j) = \theta_p + \theta_v \exp(-\theta_\beta ||I_i - I_j||_2^2), \tag{3}$$

where I_i and I_j are the colour vectors of pixels i and j respectively. θ_p, θ_v, $\theta_\beta \geq 0$ are model parameters which can be set by cross validation. Here we use the same parameters as [23] as we have a limited amount of labelled data and our street views are similar in style to theirs. The intensity-based pairwise potential at the pixel level induces smoothness of the solution encouraging neighbouring pixels take the same label.

Higher Order Potential: The higher order term $\psi_c^S(\mathbf{x}_c)$ describes potentials defined over overlapping superpixels

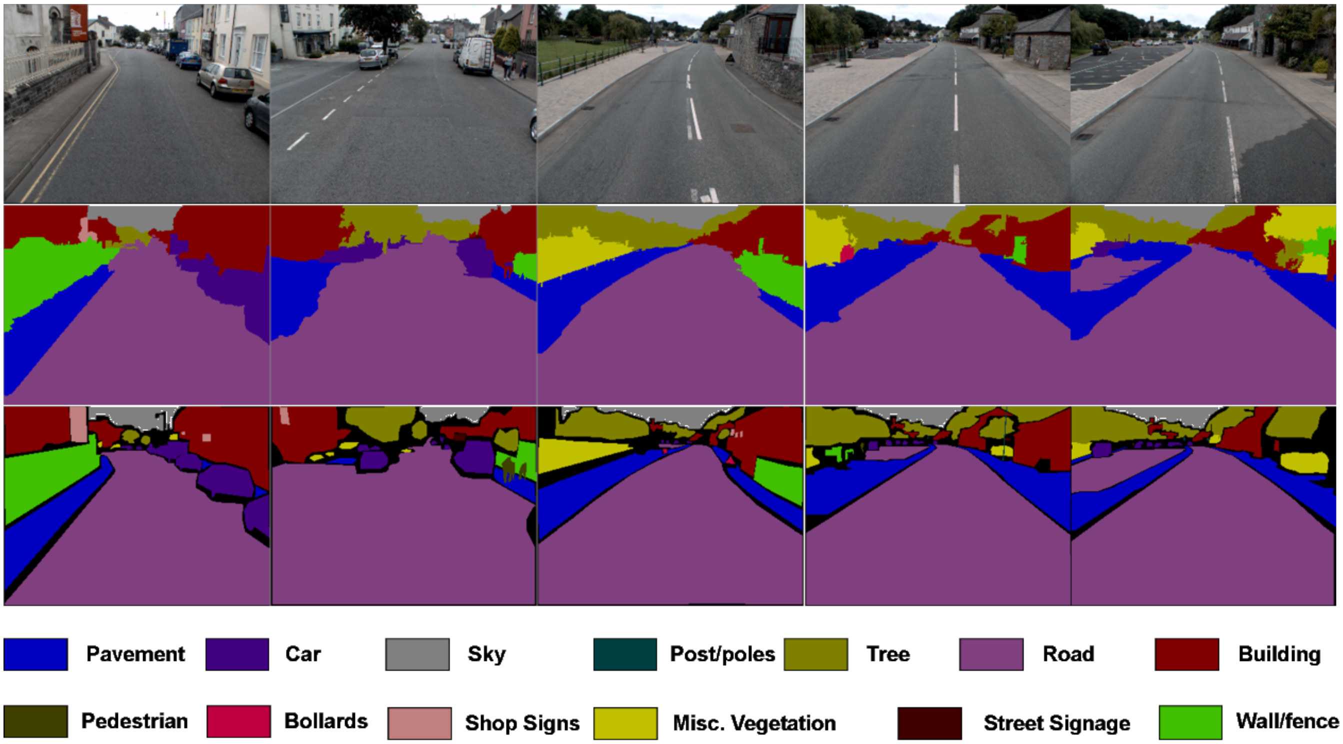

Fig. 3. *Semantic Image Segmentation*: The top row shows the input street-level images followed by the output of our first CRF model. The last row shows the corresponding ground truth for the images. Best viewed in colour.

obtained using multiple unsupervised meanshift segmentations [6]. The potential models the likelihood of pixels in a given segment taking similar label and penalizes partial inconsistency of superpixels. The costs are learnt using bag-of-words classifier. Please read [14] for more details. The energy minimization problem is solved using graph-cut based alpha expansion algorithm [2].

B. Homography

In this section we describe how the image and the ground plane are related through a homography. We use a simplifying assumption that the world comprises of single flat ground plane. Under this assumption we estimate heuristically a frontal rectangular ground plane patch for each image, using the camera viewing direction and camera height; we assume a constant camera height. For computing the camera viewing direction we refer to [12]. Each ground plane pixel z_i in the ground plane patch is back-projected into the the k^{th} image X^k, as $x_j^k = P^k z_i$, where P^k is the corresponding camera matrix. This then allows us to define the set T_i of valid ground plane to image correspondences (as shown in Fig. 2(b)). Fig. 4 shows an example ground plane patch and a registered corresponding image region. Any object such as the green pole (Fig. 4(a)), that violates the flat world assumption will thus create an artifact on the ground plane that looks like a shadow, seen in the circled area in Fig. 4(a). If we only have a single image then this shadowing effect would cause the pole to be mapped onto the ground plane incorrectly as seen in Fig. 4(b). If we have multiple views of scene then we will have multiple shadowing effects. These shadows will overlap where the violating object meets the ground, as seen in Fig. 4(c). This is precisely the part of the object we are interested while building an overhead map. Also the flat ground around the object, shown in blue, will be correctly mapped in more views. This means that if we use voting over many views we gain robustness against violations of the flat world assumption.

C. Dense Semantic Map

Our semantic map that represents the ground plane as if being viewed from above and is formulated as a pairwise CRF. The energy function for the map E^M is defined over the random variables $\mathbf{Z} = \{Z_1, Z_2, ..., Z_N\}$ corresponding to the ground plane map pixels as

$$E^M(\mathbf{z}) = \sum_{i \in \mathcal{V}} \psi_i^M(z_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}^M(z_i, z_j), \qquad (4)$$

With the same label set as that of the image domain (§II-A), except for *sky* that should not be on the ground plane.

Unary potential: The unary potential $\psi_i^M(z_i)$ gives the cost of the assignment: $Z_i = z_i$. The potential is calculated by aggregating label predictions from many semantically segmented street-level images (§II-A) given the registration (§II-B). The unary potential $\psi_i^M(z_i = l)$ is calculated as:

$$\psi_i^M(z_i = l) = -log(\frac{1 + \sum_{t \in T_i} \delta(x_t = l) c_t}{|L| + \sum_{t \in T_i} c_t}) \qquad (5)$$

where $|L|$ is the number of the labels in the label set. T_i denotes the set of image plane pixels with valid registration via the homography as defined in II-B. The factor c_t is used

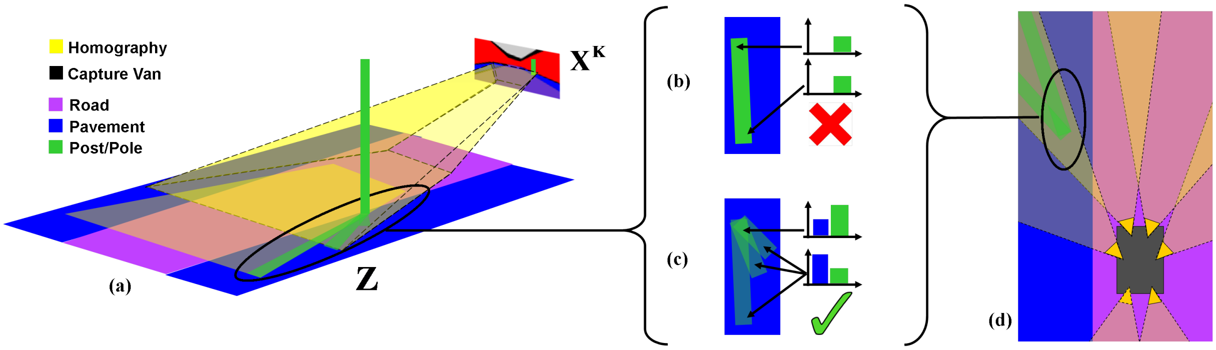

Fig. 4. *Flat World Assumption:* In order to find a mapping between the local street level images to our global map a we use a the simplifying assumption that the world is flat. The camera's rays act like a torch, depicted in yellow in (a). The green pole, which violates this assumption creates an artifact on the ground plane that looks like a shadow, seen in the circled area in (a). With single image, the shadowing effect would cause the pole to be mapped onto the ground plane incorrectly as seen in (b). With multiple views of scene, we have multiple shadowing effects. These shadows overlap where the violating object meets the ground, as seen in (c). Each view votes onto the ground plane, results in more votes for the correct label. (d) shows the effect of the multiple views from different cameras. Best viewed in colour.

to weight the effect of the image plane pixel contributing to the ground plane. Ideally we would like to have a lower weight for all the image pixels that are farther away from that camera, but due to lack of the training data we weight all the pixels uniformly. This kind unary potential represents a form of naive Bayes probabilistic measure which is easy to update in an online fashion. This voting scheme captures information across multiple views and time inducing robustness in the system.

Pairwise potential: The pairwise potential $\psi^M_{ij}(z_i, z_j)$ represents the cost of the assignment: $Z_i = z_i$ and $Z_j = z_j$ and takes the form of a Potts model:

$$\psi^M_{ij}(z_i, z_j) = \begin{cases} 0 & \text{if } z_i = z_j, \\ \gamma & \text{otherwise,} \end{cases} \tag{6}$$

The pairwise potential encourages smoothness over the ground plane. A strong hypothesis for a particular label in the ground plane position is likely to be carried over to its neighbours. Thus, for uncertain regions, this property will help in getting correct labelling. The pairwise parameter γ, is set manually by eye on unlabelled validation data.

III. Experiments

For qualitative and quantitative evaluation of our method, we create a dataset that is 14.8 km long captured by YottaDCL[2] using a specialized road vehicle with 6 mounted cameras, 2 frontal, 1 either side and 2 rear facing. The van used to capture the data is depicted in Fig. 5.

YottaDCL Pembrokeshire Dataset: Our dataset comprises of an 8000 long sequence of images captured from each camera at every 2 yards (approximately 1.8 meters) interval in the Pembrokeshire area of the United Kingdom. The images at full resolution are 1600 × 1200. This area contains urban, residential, and some rural locations making it a quite varied and challenging real world dataset. The camera parameters, computed from the GPS and odometry information are also provided. We have manually selected a small set of 86 images from this sequence for ground truth annotation with 13 object classes, 44 of the labelled images are used for training the potentials of E^S and remaining 42 are used for testing. The 13 labels are road, building, vehicle, pedestrian, pavement, tree, misc-vegetation, sky, street-bollard, shop-sign, post-pole, wall-fence, street-signage. For evaluation, we measure the global number of correctly labelled pixels over all classes on the re-projected results into the image plane using the registration process.

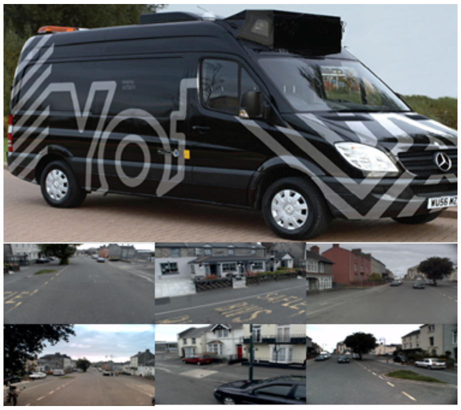

Fig. 5. *YottaDCL van:* The van used to capture the data by YottaDCL [8] with 6 mounted cameras, 2 frontal, 1 either side and 2 rear facing. Some images captured by the van are also shown.

Results: Fig. 3 shows the output of the first level of our CRF framework. The first row shows the street-level images captured by the vehicle. The second and the third row show the semantic image segmentation of the street images and the corresponding ground truth. For computational efficiency we have downsampled the original image size into 320 × 240. Fig. 6 shows the map and corresponding street imagery. The arrows shows the positions on the ground plane map where the image objects are located. We can see in Fig. 6(a) a T-junction and cars can be seen in both the map and images. Similarly in Fig. 6(b) the cars are shown in the map. In Fig. 6(c) a car-park (modelled as road) and a fence are shown in both map and the images and finally in Fig. 6(d) we see the

[2] This data has been used in the past for real-world commercial applications and has not been designed for the purpose of this paper. Please see http://www.yottadcl.com/

buildings and the pavement in the map. In the semantic map we do not have the sky as it lies above the horizon in the street images and are not captured in the ground plane-image registration. The classes like road, pavement, building, fence, vehicle, vegetation, tree which have long range contextual information and span across many images, appear more frequently in the map. For the similar reason, the smaller objects classes like pedestrian, post-pole, street-signage do not show up often. This is because the images are taken at approximately every two yards, so the objects without a long range contextual information tend not to appear in the map. Quantitatively, we achieve a global pixel accuracy of 82.9%. Fig. 7 shows another output of our algorithm, where the map spans a distance of 7.3km. Both the vehicle track (white track on Google Earth) and the semantic map output (overlaid image) is shown. The map building for this example takes 4000 consecutive frontal street-level images as the input.

IV. Conclusions and Future Work

We have presented a framework where we generate a semantically labelled overhead view of an urban region from a sequence of street-level imagery. We formulated the problem using two conditional random fields. The first one performs semantic image segmentation locally at the street level and the second updates a global semantic map. We have demonstrated results for a track of 7.3 km showing object labelling of the map.

In the future we aim to scale the work to much larger regions, and even the whole of the U.K. This poses a challenge of accommodating the variety of visual data for classification, taken across geographical locations spanning thousands of km's of roadways. We would like to go beyond the current flat world assumption and create a dense semantic reconstruction using the multiple street view images. We also aim to increase the computational efficiency of the system. For this we would like to speedup the feature calculation using GPU or FPGA implementation of the same and use computationally efficient binary features like [3], [16]. The inference stage can be made faster using dynamic approaches for MAP solutions [13] or the more recent fast mean-field based approach [24]. We hope that these measures will improve the overall system time to deploy it in a real vehicle.

Acknowledgements

We thank Jonathan Warrell, Vibhav Vineet and Julien Valentin for useful discussion in this paper. This work is supported by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. P. H. S. Torr is in receipt of Royal Society Wolfson Research Merit Award.

References

[1] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, pages I: 105–112, 2001.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23:2001, 2001.

[3] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: binary robust independent elementary features. In *ECCV*, ECCV'10, pages 778–792, Berlin, Heidelberg, 2010. Springer-Verlag.

[4] A. Chambers, S. Achar, S. Nuske, J. Rehder, B. Kitt, L. Chamberlain, J. Haines, S. Scherer, and S. Singh. Perception for a river mapping robot. In *IROS*, pages 227–234. IEEE, 2011.

[5] J. Civera, D. Galvez-Lopez, L. Riazuelo, J. D. Tardos, and J. M. M. Montiel. Towards semantic slam using a monocular camera. In *IROS*, pages 1277 –1284, sept. 2011.

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[7] H. Dahlkamp, G. Bradski, A. Kaehler, D. Stavens, and S. Thrun. Self-supervised monocular road detection in desert terrain. In *RSS*, Philadelphia, 2006.

[8] Y. DCL. Yotta dcl case studies, retrived April 2010.

[9] B. Douillard, D. Fox, F. Ramos, and H. Durrant-Whyte. Classification and semantic mapping of urban environments. *IJRR*, 30(1):5–32, Jan. 2011.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[11] Google Earth. Google earth (version 6.0.0.1735 (beta)) [software]. mountain view, ca: Google inc.

[12] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[13] P. Kohli and P. H. S. Torr. Dynamic graph cuts for efficient inference in markov random fields. *IEEE PAMI*, 29(12):2079–2088, Dec. 2007.

[14] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

[15] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. In *BMVC*, 2010.

[16] S. Leutenegger, M. Chli, and R. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *ICCV*, 2011.

[17] A. Milella, G. Reina, J. P. Underwood, and B. Douillard. Combining radar and vision for self-supervised ground segmentation in outdoor environments. In *IROS*, pages 255–260, 2011.

[18] P. Newman, G. Sibley, M. Smith, M. Cummins, A. Harrison, C. Mei, I. Posner, R. Shade, D. Schroeter, L. Murphy, W. Churchill, D. Cole, and I. Reid. Navigating, recognizing and describing urban spaces with vision and lasers. *IJRR*, 28(11-12):1406–1433, Nov. 2009.

[19] A. Nüchter and J. Hertzberg. Towards semantic maps for mobile robots. *Robot. Auton. Syst.*, 56(11):915–926, Nov. 2008.

[20] I. Posner, M. Cummins, and P. Newman. Fast probabilistic labeling of city maps. In *RSS*, Zurich, Switzerland, June 2008.

[21] I. Posner, M. Cummins, and P. M. Newman. A generative framework for fast urban labeling using spatial and temporal context. *Auton. Robots*, 26(2-3):153–170, 2009.

[22] J. Shotton, J. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, pages 1–15, 2006.

[23] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009.

[24] V. Vineet, J. Warrell, and P. H. S. Torr. Filter-based mean-field inference for random fields with higher order terms and product label-spaces. In *ECCV*, pages 1–10, 2012.

[25] K. M. Wurm, R. Kümmerle, C. Stachniss, and W. Burgard. Improving robot navigation in structured outdoor environments by identifying vegetation from laser data. In *IROS*, pages 1217–1222, Piscataway, NJ, USA, 2009. IEEE Press.

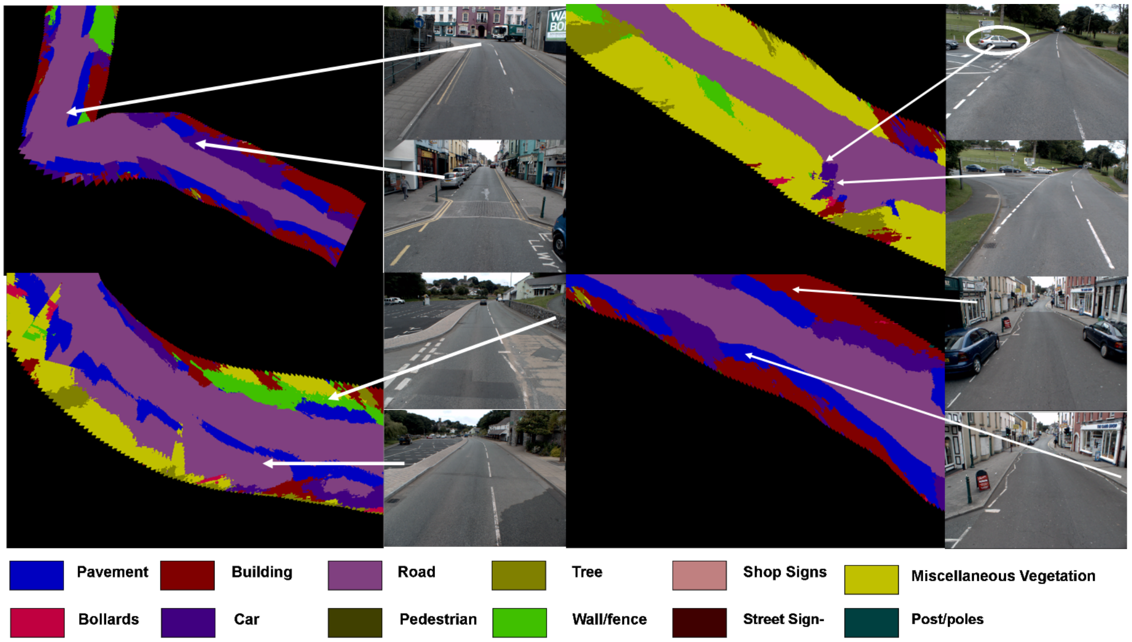

Fig. 6. Overhead Map with some associated images. The arrow shows the positions in the map where the image were taken. In (a) we see from the image there is a T-junction, which is also depicted in the map. The circle in the image (b) shows the car in the image, which has been labelled correctly in the map. Similarly in (c) the fence and the carpark from the images are also found in the map. In (d) arrows showing the pavement and the buildings being mapped correctly. Best viewed in colour.

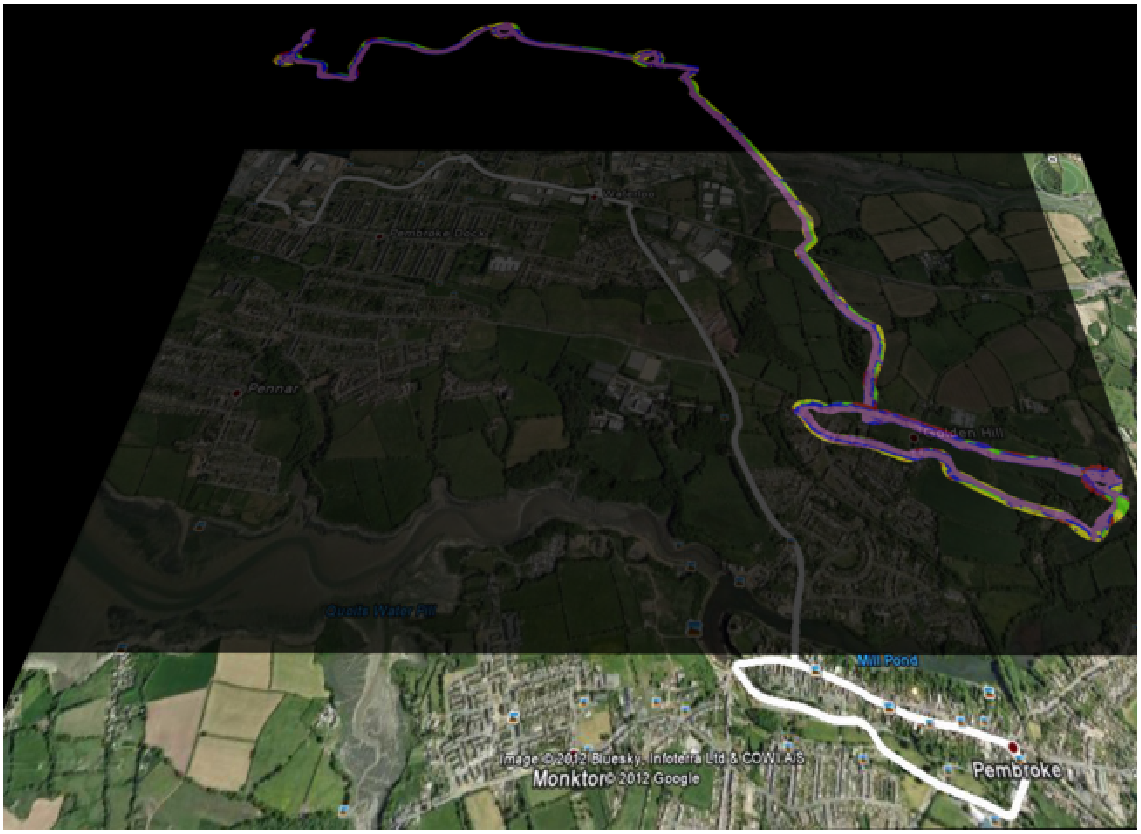

Fig. 7. *Map output 2* Semantic Map showing a road path of 7.3km in the Pembroke city, UK. The Google Earth image (shown only for visualization purpose) shows the actual path of the vehicle (in white). The top overlaid image shows the semantic map output corresponding to the vehicle path. Best viewed in colour.