Mesh Based Semantic Modelling for Indoor and Outdoor Scenes

Julien P. C. Valentin[1,3] Sunando Sengupta[1,3] Jonathan Warrell[1] Ali Shahrokni[2] Philip H. S. Torr[1]
{julien.valentin-2010, ssengupta, jwarrell, philiptorr}@brookes.ac.uk,
ali.Shahrokni@2d3sensing.co.uk
[1]Oxford Brookes University [2]2d3 Ltd. *

Abstract

Semantic reconstruction of a scene is important for a variety of applications such as 3D modelling, object recognition and autonomous robotic navigation. However, most object labelling methods work in the image domain and fail to capture the information present in 3D space. In this work we propose a principled way to generate object labelling in 3D. Our method builds a triangulated meshed representation of the scene from multiple depth estimates. We then define a CRF *over this mesh, which is able to capture the consistency of geometric properties of the objects present in the scene. In this framework, we are able to generate object hypotheses by combining information from multiple sources: geometric properties (from the 3D mesh), and appearance properties (from images). We demonstrate the robustness of our framework in both indoor and outdoor scenes. For indoor scenes we created an augmented version of the* NYU *indoor scene dataset (*RGB-D *images) with object labelled meshes for training and evaluation. For outdoor scenes, we created ground truth object labellings for the* KITTI *odometry dataset (stereo image sequence). We observe a significant speed-up in the inference stage by performing labelling on the mesh, and additionally achieve higher accuracies.*

1. Introduction

In this paper we propose a method to generate a semantically labelled reconstruction of any scene. In our approach, the scene is reconstructed in the form of a mesh, computed from a sequence of depth estimates, and is annotated with semantic object labels. This provides a more consistent approach to scene segmentation compared with independent labelling of a sequence of images. Moreover, by virtue of working on meshes, our method is *highly efficient* in the inference stage. This form of semantically annotated 3D representation is necessary to allow robotic platforms to understand, interact and navigate in a structured indoor environment [2, 12, 26] or outdoor scenes [1, 30, 7]. Our labelling framework *is designed for both types of scenes*. For semantic indoor scene reconstruction we use images from RGB-D sensors, and for the outdoor road scenes, we use a sequence of stereo images.

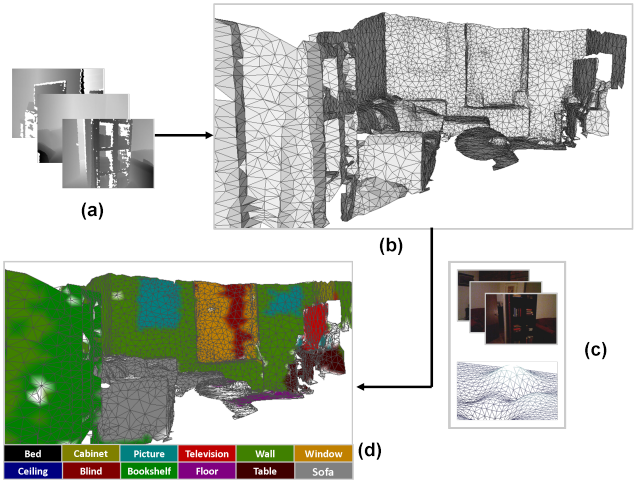

Figure 1. Semantic Mesh Segmentation: (a) shows the input to our system which is a sequence of images with depth estimates (here we show the images of a living room captured using a RGBD camera [24]), (b) the depth estimates are used to generate a mesh based representation of the scene using the method of [31], (c) the system combines image level information and geometric contextual information from the mesh to perform object class labelling, (d) object class labelling is performed by establishing a conditional random field (CRF) on the mesh, with local neighbourhood interactions defined by the neighbouring faces in the mesh. (Best viewed in colour)

The problem of semantic object labelling has been studied extensively and has seen major improvements [16, 23, 29]. However, most of these algorithms work in the image domain, where each pixel in the image is classified with an object label. The image data is intrinsically incomplete as it consists of a projection from a 3D world into a 2D plane. As a result, crucial data such as shape, geometric layout and the scale of objects, which are strong cues for object class segmentation, are lost. Moreover, the independent treatment of each image of a scene results in a series of inconsistent

*[3] The authors assert equal contribution and joint first authorship.

object-class labellings.

The advent of inexpensive depth sensors has significantly encouraged developments in scene reconstruction using streams of depth and RGB data [18, 19]. Large-scale stereo image sequences have also been used to generate dense reconstructions of urban road scenes [10]. In most cases, the scene is reconstructed as a mesh [19]. Segmentation of meshes has been a subject of much research in computer graphics [14, 5, 13]. However, most of these methods consider only geometric properties, ignoring the appearance.

Recently, [15] proposed combining both visual and geometric cues for labelling indoor scene point clouds, but they use a restricted learning framework and an inefficient (slow) inference method. Furthermore, the method uses geometric clustering to establish pairwise connections which can produce inconsistency along object boundaries. Similarly, an attempt to label indoor scene images captured using RGB-D sensor has been made in [24] where a classifier was trained in the image domain, along with a 3D distance prior to aid their classification. This was extended in [25] where object surface and support relations of an indoor scene were used as a prior to perform indoor scene segmentation. However both of these methods ignore the full geometric properties of the objects in the scene. For outdoor scene labelling, most of the work has concentrated on classification in the image domain [27, 4, 17] or using a coarse level 3D interpretation [8]. In [22] a semantic 3D reconstruction is generated but the object labelling is performed in the image domain, and then projected to the model. As a result, these methods fail to take advantage of the structured geometry of the road scenes.

In this work, we tackle the problem of semantic scene reconstruction (for both indoor and outdoor scenes) in 3D space by combining both structural and appearance cues. Our approach to semantic mesh segmentation is illustrated in Fig. 1. The input to our method is a sequence of images and their depth estimates. The depth estimates are used to generate a triangulated mesh representation of the scene [20, 31], enabling us to capture its inherent geometry. We propose a cascaded classifier to learn the geometric cues from the mesh, and appearance cues from images in an efficient learning framework [23]. For this purpose, we augment the NYU depth dataset [24] with semantically annotated ground truth meshes for training and evaluation purposes. Furthermore we solve the labelling problem in 3D by defining a conditional random field (CRF) over the scene mesh, effectively exploiting the scene geometry. As a result, we achieve a significant speedup at the inference stage (20× for outdoor and 1000× for indoor scenes) in comparison to the methods of [24, 15, 16]. For outdoor scene labelling, we use the stereo image sequence of the KITTI odometry dataset [9], for which we generated ground truth object class labellings for training and evaluation purposes.

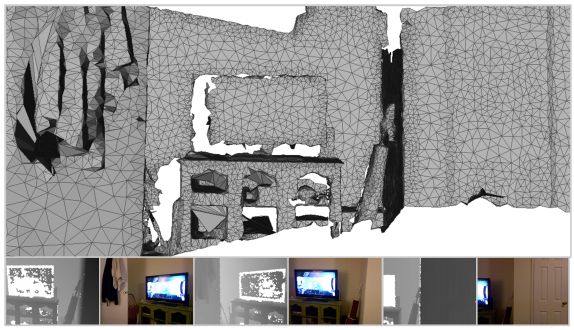

Figure 2. Reconstruction of a living room from the NYU dataset. The bottom row shows the depth and the color images of the scene. The depth images are fused into a single globally consistent scene representation (top row).

In summary, our contributions are:

- The formulation of the labelling problem in the 3D space resulting in a significant speedup in the inference stage: §2.
- A cascaded boosting approach to train our classifiers, combining both geometric and image based appearance cues for object segmentation: §3.
- For indoor scenes, the augmentation of the NYU dataset with ground truth meshes to facilitate learning on the mesh. For outdoor scenes, the creation of a per-pixel object class labelling for the KITTI dataset[1]: §4.

2. Semantic labelling of the mesh based scene representation

To generate a semantically annotated meshed representation of a scene, we use a sequence of depth estimates along with images describing the scene. They are merged using the Truncated Signed Distance Function (TSDF) and used to generate a mesh. Geometric features are computed on the mesh directly; visual features are computed on the images and then projected to the corresponding faces of the mesh. A graph homomorphic to the mesh is then built with the appropriate potentials. Finally, an approximate MAP inference on the CRF is solved which gives us the labelling of the mesh. In the remainder of the section, the above steps are explained in more detail.

2.1. Mesh Estimation

To estimate a meshed representation of the scene we use a sequence of depth estimates which are obtained from a depth sensor (providing a stream of RGB-D images) for indoor scenes and stereo image pairs for outdoor scenes. The depth estimates are incrementally fused into a single 3D reconstruction using the volumetric TSDF representation [6].

[1] project website: http://cms.brookes.ac.uk/research/visiongroup/projects.php

A signed distance function assigns to each voxel a value equal to the signed distance to the closest surface interface (zero crossing), with positive increasing values corresponding to free space and negative decreasing values corresponding to points beneath the surface. The representation allows for the efficient combination of multiple noisy surface measurements, obtained from different depth maps by averaging signed the distance measures from every depth map at each voxel. This smoothes out the irregularities in the surface normals of the individual depth maps. For further details we refer the reader to [18, 19]. In order to obtain a complete meshed surface, we first infer an iso-surface from the TSDF field by finding all the zero crossings. A triangulated mesh corresponding to the zero valued iso-surface is extracted using [20, 31]. Fig. 2 shows an example output of surface reconstruction along with the images of the scene.

2.2. CRF energy model

We use a Conditional Random Field (CRF) based approach, defined over a mesh structure, to perform the semantic labelling of the mesh. Consider a set of random variables $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$, where each variable $X_i \in \mathcal{X}$ takes a value from a pre-defined label set $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$. A labelling $\mathbf{x}$ refers to any possible assignment of labels to the random variables and takes values from the set $\mathcal{L}^N$. The CRF is defined over the estimated mesh (given by a set of face indices $\mathcal{F} = \{1, 2, \ldots, N\}$), where each mesh face, $i \in \mathcal{F}$ is associated with its corresponding random variable X_i. This representation allows us to associate the geometric properties of the face with the random variable X_i. Each mesh face i is registered to a set of images through the camera projection matrix, giving a set $\tau_i \subset \mathcal{P}$ of image pixels for each mesh face where $\mathcal{P}$ is the set of all the image pixels. Let $\mathcal{N}$ be the neighbourhood system of the random field defined by sets $\mathcal{N}_i, \forall i \in \mathcal{F}$, where $\mathcal{N}_i$ denotes the set of all the neighbours (in this case adjacent faces) of the variable X_i. The corresponding Gibbs energy of the mesh is given as:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{F}} \psi_i(x_i) + \sum_{i \in \mathcal{F}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j) \qquad (1)$$

The most probable or maximum a posteriori labelling $\mathbf{x}^*$ of the CRF corresponds to the minimum energy of the graph. The energy minimisation problem is solved using the graph-cut based alpha expansion algorithm [3], giving us the labelling of the mesh.

Unary potential The unary potential ψ_i describes the cost of a mesh face $\mathcal{F}_i$ taking a particular object class label. The form of the unary potential is the negative logarithm of the normalised output of a boosted classifier. The input to the classifier is a feature vector composed of image, geometric and contextual features. For the image level features, we use the multi-features variant [16] of the TextonBoost algorithm [23]. The geometric features used in our experiments are surface curvature, singular values extracted from principal component analysis (PCA) of local shape, shape diameter feature (SDF), shape contexts (SC) and spin images. For further details we refer to the paper [14]. Both the image and geometric features are trained together to produce a contextual feature set.

Pairwise Potential The pairwise potential takes the form of a contrast-sensitive Potts model:

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

where the function $g(i, j)$ is an edge feature based on the difference in colour and geometry of neighbouring faces:

$$g(i,j) = \theta_1 + \theta_2 \exp(-\alpha\omega(i,j) - \frac{\sum_{k \in \tau_i, l \in \tau_j} d(k,l)}{|\tau_i|}) \qquad (3)$$

where $\omega(i, j)$ is the measure of the exterior dihedral angle between face i and j and $d(k, l)$ is the Euclidian distance between the colour values at pixel k and l. The term $d(k, l) = 0$ if the pixels k and l do not belong to the same image while τ_i and τ_j denote the set of image pixels registered to the faces i and j respectively.

3. Cascaded learning

There has been extensive work on image-level segmentation and promising recent advances on geometry-level segmentation. With the advent of cheap RGBD sensors, there is a need for a flexible and efficient framework to use and combine evidence from both domains simultaneously. Our approach initially treats features from the image and the geometry separately, and then allows them to be combined along with contextual features by repeated application of the JointBoost algorithm [28]. In the following, we use the shorthand x_i to denote the label of the ith element of the domain (pixel for images, faces for the mesh) on which the training is performed, and $\mathbf{z}_i$ as the feature vector associated with x_i. Each step of our method learns a classifier based on the current set of features $\mathbf{z}$. At the end of each step, to this set of features is added a feature extracted from the current estimate of the distribution of labels around each element x_i, which is estimated using the JointBoost classifier output. The process can be repeated and is summarized in algorithm 1 (we use $S = 3$) .

From the images, we train a TextonBoost [23] classifier provided by visual words. From that step, we project each pixel onto the mesh. Given the set of pixels that land on a specific face:

1. Compute the posterior distribution of each of these pixels using TextonBoost's classifier and then average them.

2. Build a cumulative normalized histogram of visual words that land on that face.

From these two steps we form the feature set $\boldsymbol{z}^I$.

In addition to the image fatures, for each face we also use 3D features using the procedure described in [14]. From that step, we extract geometric features and geometric contextual features. That set is referred to as $\boldsymbol{z}^M$, and the set formed by $\boldsymbol{z}^I$ and $\boldsymbol{z}^M$ to as $\boldsymbol{z}$.

Finally, we build a set of graphs $\mathcal{G}$ to define several neighbourhoods for each x_i. Each graph is built such that each node is connected to every other node that is within a specific geodesic distance. We are then able to optimise the learning on the mesh using that set of graphs and $\boldsymbol{z}$ as input into algorithm 1.

The unary potential from our energy model corresponds to the classifer $\phi(\boldsymbol{z}_i, l)$ extracted from algorithm 1, to which a softmax was applied. Given the unary potential, we can set the model parameters θ_1, θ_2 and α of the pairwise potential by cross validation.

Algorithm 1: Training procedure

Input: A generic set of features $\boldsymbol{z}$ together with the ground-truth in the selected domain, and a set of graphs defining different sets of neighbours each x_i

Output: A multi label classifier score $\phi(\boldsymbol{z}_i, l)$

Algorithm:

Learn an initial classifier $\phi^0(\boldsymbol{z}_i, l)$ using the JointBoost algorithm provided by $\boldsymbol{z}$ and the ground truth

for $s = 1$ *to* S **do**

  For each x_i, extract the posterior distribution for each label l as $\frac{\exp(\phi^s(\boldsymbol{z}_i,l))}{\sum_{k\in L}\exp(\phi^s(\boldsymbol{z}_i,k))}$

  foreach $g \in \mathcal{G}$ **do**

    Using the neighbourhood defined in g, compute the average posterior distribution in the neighbourhood of x_i. This forms a new set of features $\boldsymbol{c}$.

    $\boldsymbol{z}_i := [\boldsymbol{z}_i; \boldsymbol{c}]$

  end

  Learn $\phi^s(\boldsymbol{z}_i, l)$ using the JointBoost algorithm provided by the extended feature set and the ground truth

end

return $\phi^s(\boldsymbol{z}_i, l)$

4. Experiments

To demonstrate the effectiveness of our proposed system, we have evaluated it using indoor scenes on the NYU depth dataset [24] augmented with meshes, and for outdoor urban road scenes, we used the publicly available KITTI dataset [9].

4.1. Reconstruction of indoor scene

Dataset The NYU dataset [24] comprises indoor scene video sequences captured using Microsoft Kinect. Ground truth annotations are provided for 12 object classes. The class labels are *bed, cabinet, picture, television, blind, ceiling, sofa, wall, bookshelf, floor, table and window*. We enhance the NYU indoor dataset [24] with labelled meshes for a representative subset of the total number of scenes, which we use for training and evaluation purposes. Meshes are estimated using the method described in §2.1. To label each face of the mesh, we first build an histogram that will be used to count the number of votes each label gets. We then project each ground truth pixel onto the corresponding face in the mesh and increase the vote count for the label associated with that pixel by one. Finally, we pick the label that has accumulated the most votes. Some of the scenes have not been considered, as the cameras were too close to the wall, leading to erroneous ICP registration and incoherent 3D reconstruction with holes in the mesh. In our dataset, we selected 33 out of the 64 scenes of the NYU depth dataset. Our dataset comprises bedroom, bathroom, living room, kitchen and office scenes. Each of these meshes have associated ground truth labelled images. We use 23 mesh scenes (370 ground truth images) for training and 10 (221 ground truth images) for testing, and train two stages for each cascade.

Fig. 3 shows the qualitative results on the indoor scenes from the NYU dataset [24], where the semantically segmented mesh is shown along with the corresponding images. Fig. 3(a) shows an example of an office scene. The classes such as floor, ceiling and walls which have a strong horizontal geometric orientation have been classified correctly. In the scene, we also see that the bookshelf is classified as cabinet, which has similar geometric and appearance properties. Fig. 3(b) shows the semantic mesh corresponding to a living room. Here, we can see the cabinet properly classified in the mesh, as shown by the arrow in the image. However, the television is misclassified as a picture, with which it has strong similarities in terms of geometric structure. Finally in Fig. 3(c) a kitchen scene is shown where the kitchen cabinet and the window are correctly classified in the mesh. The experiments show the strong influence of learning the contextual features.

Table 1 compares the accuracy of our method with previous approaches, where we report the percentage of correctly labelled mesh faces. We compare our method with the results of [16], evaluating the set of the images corresponding to our augmented dataset as described in 4.1. To compare with [16], we used the publicly available ALE library, with

unary, pairwise and higher order terms added in the energy function. We also note that for our method, we report scores obtained in the mesh domain only as noisy camera estimations lead to inferior results when back-projecting labelling results from the mesh domain to the image domain. The method of [24] achieves a mean recall score of 56.5%, averaged over 10 training/testing splits. However, their result is evaluated on the entire dataset, and so it is not directly comparable to ours. Furthermore, we are comparing the labelling of meshes rather than using images as is done in [24]. We also test the effect of the contextual features in our method for indoor scenes. The learning procedure selects the contribution of each visual and geometric feature. Table 3 shows the selection percentage for each of these features at the end of the two stages of cascaded learning (C_3). The change in the level of contribution of geometric features relative to the image based feature is noticeable in both stages of C_3. In the first stage, the contribution of image level unary features is around 54%, but in the second stage of C_3, it is reduced to 35% and the share of geometric and contextual features is increased. The results from table 1 suggest that the proposed learning method and the accumulation of evidence of the same scene from multiple angles leads to very strong results using auto-context only.

Timing wise, inference on meshes is computationally more efficient than inference on images as we do not have to perform a per-pixel labelling for all the images describing the scene. As [24, 16], inference is performed using the graph cut based alpha expansion method of [3]. Our test set comprises 10 meshed scenes with 221 associated ground truth object labelled images, each of size 640×480. Overall, the reconstructed meshed models have $78,340$ vertices and $141,870$ faces. This leads to an approximate 1000x speedup in the inference stage (see Table 2). Our method performs the meshing of the scene prior to the inference stage, using the method of [19]. We compute the timings on a single core 2.66Ghz Intel Xeon processor and take the average of three runs.

[16]	Our cascaded learning	Our full model
77.05%	78.2%	78.5%

Table 1. Quantitative Evaluation for NYU dataset. The figure shows the percentage of the correctly classified mesh triangles. The second column presents our results with the unary potential only and the third column correspons to the results we get with both the unary and pairwise potentials.

Image Level Inference [16, 24]	Ours
1098.76s	1.057s

Table 2. Inference times for Indoor Scene segmentation (in seconds).

4.2. Reconstruction of outdoor scene

Outdoor dataset For the evaluation of our method on outdoor scenes, we use the KITTI odometry dataset [9]. Stereo images are captured using a specialised car in urban, residential and highway locations, making it a varied and challenging real world dataset. We have manually annotated a set of 45 images for training and 25 for testing with per-pixel class labels. The class labels are road, building, vehicle, pavement, tree, sky, signage, post/pole, wall/fence. We intend to release these hand labelled ground truth images. Due to lack of sufficient training data, we do not have ground truth meshes to learn geometric cues. Hence we use only the appearance cues from images and a single cascade to perform semantic mesh labelling.

For reconstruction from a stereo image sequence, we need to estimate the camera poses. This comprises two main steps, namely feature matching and bundle adjustment. We assume calibrated cameras are positioned on a rigid rig. To obtain the feature correspondences, we perform both stereo matching (between left and right pairs) and frame by frame matching (consecutive images for both left camera and right camera). Once the matches are computed, the corresponding feature tracks are generated. All the stereo matches and the corresponding frame-to-frame matches are kept in the track. As the road scene images are subject to high glare, reflection and strong shadows, it is important to generate good feature matches for accurate camera estimation. Having this agreement between both the stereo and ego motion helps bundle adjustment to estimate the camera poses and feature points more accurately. Given the feature track database, the camera poses and the associated feature points are estimated using a Levenberg-Marquardt optimiser. As we are dealing with street-level sequences of arbitrary length, a global optimisation is computationally infeasible, and may be unnecessary, since only the structure of the world near to the current frame is of interest in many applications. Hence, we use only the last 20 frames for the bundle adjustment which we found to be a good compromise between speed and accuracy. The estimated camera poses are used to fuse the depth estimates (computed from stereo disparity) in an approach similar to [19]. Finally, surface reconstruction is performed using the marching tetrahedrons algorithm [20] which extracts a triangulated mesh corresponding to the zero valued iso-surface.

Fig. 4 demonstrates the qualitative results of our approach for the outdoor sequence from the KITTI dataset [9]. The scene is reconstructed from 150 stereo image pairs taken from the moving vehicle. The reconstructed model in Fig. 4 comprises 703K vertices and 1.4 million faces. Visually, we can see the accuracy of the reconstructed semantic model. A close up view of the semantic model is shown in Fig.5. The reconstructed model captures fine details which are evident on the pavements, cars, road, fence etc. The ar-

Features	Curvature	PCA	SC	GCF	SPIN	SDF	HVW	2D unaries
Stage 1	2%	2%	4%	34%	1%	1%	2%	54%

Features	Curvature	PCA	SC	GCF	Contextual Features	SDF	HVW	2D unaries
Stage 2	1.33%	0.66%	4.66%	16%	38%	0.66%	3.33%	35%

Table 3. Feature contribution during the stages of Cascaded learning of C_3. (SC=Shape Context, GCF=Geometric contextual features and HVW=Histogram of visual words).

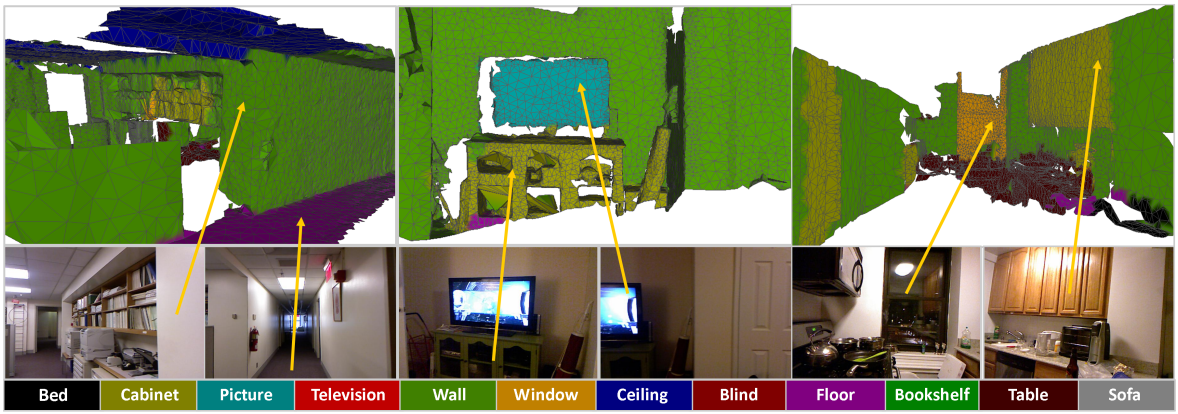

Figure 3. Semantic mesh labelling for an indoor scene [24] along with the images corresponding to the scene. (a) Example of an office scene. The arrows show the floor and the walls in both the images and output mesh. (b) Living room sequence: the scene shows the bookshelf (correctly classified) and television that is misclassified as picture. (c) Kitchen sequence: the arrow shows the cabinet and the window captured in the image and also inferred on the mesh.

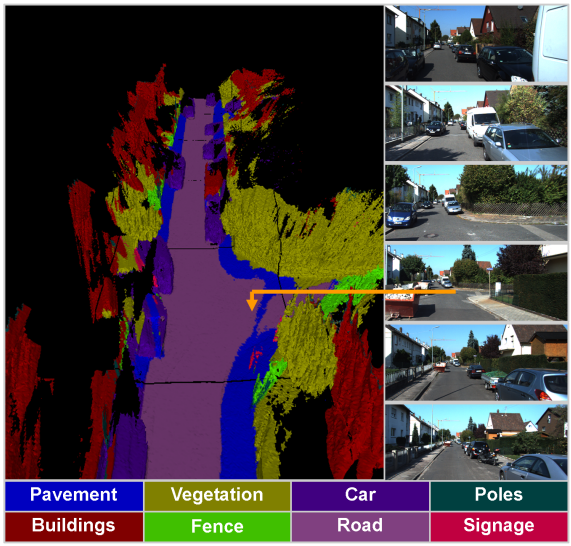

Figure 4. Reconstruction of an urban sequence (right) sequence from KITTI dataset [9]. The right column shows the sequence of stereo images that are used for reconstruction purposes. The vehicle traverses around 200 meters capturing 150 stereo image pairs. The arrow shows the bend in the road in both the image and the model (best viewed in colour).

rows relate positions in the 3D model and the corresponding images. Fig. 5(a) shows the fence, hedge and the pole (encircled) in the image and the reconstructed model. Fig. 4(b) shows the right side of the street in the reconstructed model, showing the car, pavement and the trees in both the images and the model. In all the above cases, we can see the effectiveness of our method in handling large scale stereo data sequence to generate a semantic model of the scene.

The quantitative evaluation of our reconstructed semantic model is summarised in Table 4. Evaluation is performed by projecting the ground truth labels into the mesh using the estimated camera poses, and taking a vote on the maximally occurring ground truth label for a particular mesh location. 'Global' refers to the overall percentage of the mesh correctly classified, and 'Average' is the average of the per class measures. We compare with the publicly available ALE library [16] with unary and pairwise added in the model for image level segmentation. The results from [16] are projected onto the mesh in a similar fashion to the ground truth labels. We observe an overall increase in the measures of global accuracy, average recall and average intersection-union scores. We observe an increase in performance for most of the classes in each of these measures. This can be attributed to the fact that the pairwise connections in the 3D mesh respect the structure of the objects, while in the image domain, the connections between adjacent pixels might violate the occlusion boundaries.

We also evaluate the timing performance for the inference stage for outdoor scene reconstruction. We compare our mesh based inference with the image level inference of

Method	Building	Vegetation	Car	Road	wall/fence	Pavement	Signage	Pots/Pole	**Average**	**Global**
Recall measure										
Image Level Segmentation (ALE [16])	**97.2**	84.5	**77.9**	96.2	36.6	75.2	**24.5**	31.7	65.4	83.9
Ours	96.4	**85.4**	76.8	**96.9**	**42.7**	**78.5**	24.3	**39.3**	**67.5**	**84.5**
Intersection *vs* Union										
Image Level Segmentation (ALE [16])	81.6	72.3	65.3	90.5	34.7	61.3	**19.8**	**29.1**	56.8	
Ours	**82.1**	**73.4**	**67.2**	**91.5**	**40.6**	**62.1**	16.7	25.9	**57.4**	

Table 4. Semantic Evaluation for outdoor scene KITTI dataset [9]

ALE (Image Level) [16]	Ours
1177.1±33s	60.3±2s

Table 5. Inference Timing for outdoor scenes (in seconds).

[16] with their unary and pairwise potentials. Our scene is reconstructed from 150 images, the size of each image being 1281×376. As we are trying to reconstruct an outdoor scene which spans hundreds of meters, our reconstructed mesh has around 704K vertices and 1.27 million faces. In comparison to [16], we observe a significant speedup ($20\times$) at the inference stage when performed on the mesh (see Table 5). It is worth noting that our method needs to estimate the mesh to perform the inference. However that stage can be sped up by estimating camera poses using [11] and then TSDF fusion on a GPU [21].

5. Conclusions and Future work

We have presented an efficient framework to perform 3D semantic modelling applicable to both indoor and outdoor scenes. We formulated this problem in 3D space, thereby capturing the inherent geometric properties of the scene. Further, we proposed a cascaded training framework to combine information from multiple sources: geometric properties (from 3D mesh) and appearance properties (from images). To facilitate the training/evaluation of our model, we have augmented the NYU indoor scene datasets with ground truth labelled meshes and KITTI outdoor scene dataset with object class labelling. The augmented datasets will be useful for understanding both indoor and outdoor scenes. These will be made publicly available. Finally we demonstrate substantial improvement in the inference speed (20-1000$\times$) and achieve higher accuracy for both indoor and outdoor scenes.

The current approach cannot handle large independent motion of objects, which we would like to address in the future. We would like to incorporate higher order terms in the CRF formulation, forcing similar and neighbouring triangles in the mesh to take the same label and investigate the joint optimisation of object labels and the 3D structure of the scene.

References

[1] T. Bailey and H. Durrant-Whyte. Simultaneous localization and mapping (slam): Part ii. *Robotics & Automation Magazine, IEEE*, 13(3):108–117, 2006. 1

[2] J. Biswas and M. Veloso. Depth camera based indoor mobile robot localization and navigation. pages 1697–1702. ICRA, 2012. 1

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23:2001, 2001. 3, 5

[4] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*, pages 44–57, 2008. 2

[5] X. Chen, A. Golovinskiy, and T. Funkhouser. A benchmark for 3D mesh segmentation. 28(3), Aug. 2009. 2

[6] B. Curless and M. Levoy. A volumetric method for building complex models from range images. pages 303–312. ACM, 1996. 2

[7] H. Dahlkamp, G. Bradski, A. Kaehler, D. Stavens, and S. Thrun. Self-supervised monocular road detection in desert terrain. In *RSS*, Philadelphia, 2006. 1

[8] F. Erbs, U. Franke, and B. Schwarz. Stixmentation - probabilistic stixel based traffic scene labeling. In *BMVC*, 2012. 2

[9] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, June 2012. 2, 4, 5, 6, 7

[10] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium*, Baden-Baden, Germany, June 2011. 2

[11] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE IV*, pages 963 –968, june 2011. 7

[12] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *IJRR*, 31(5):647–663, Apr. 2012. 1

[13] Q. Huang, M. Wicke, B. Adams, and L. Guibas. Shape decomposition using modal analysis. *Computer Graphics Forum*, 28(2):407–416, 2009. 2

[14] E. Kalogerakis, A. Hertzmann, and K. Singh. Learning 3d mesh segmentation and labeling, 2010. SIGGRAPH. 2, 3, 4

[15] H. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes, 2011. NIPS. 2

[16] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 1, 2, 3, 4, 5, 6, 7

[17] 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. 2

[18] R. Newcombe, S. Lovegrove, and A. Davison. Dtam: Dense tracking and mapping in real-time. In *ICCV*, 2011. 2, 3

[19] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE ISMAR*, 2011. 2, 3, 5

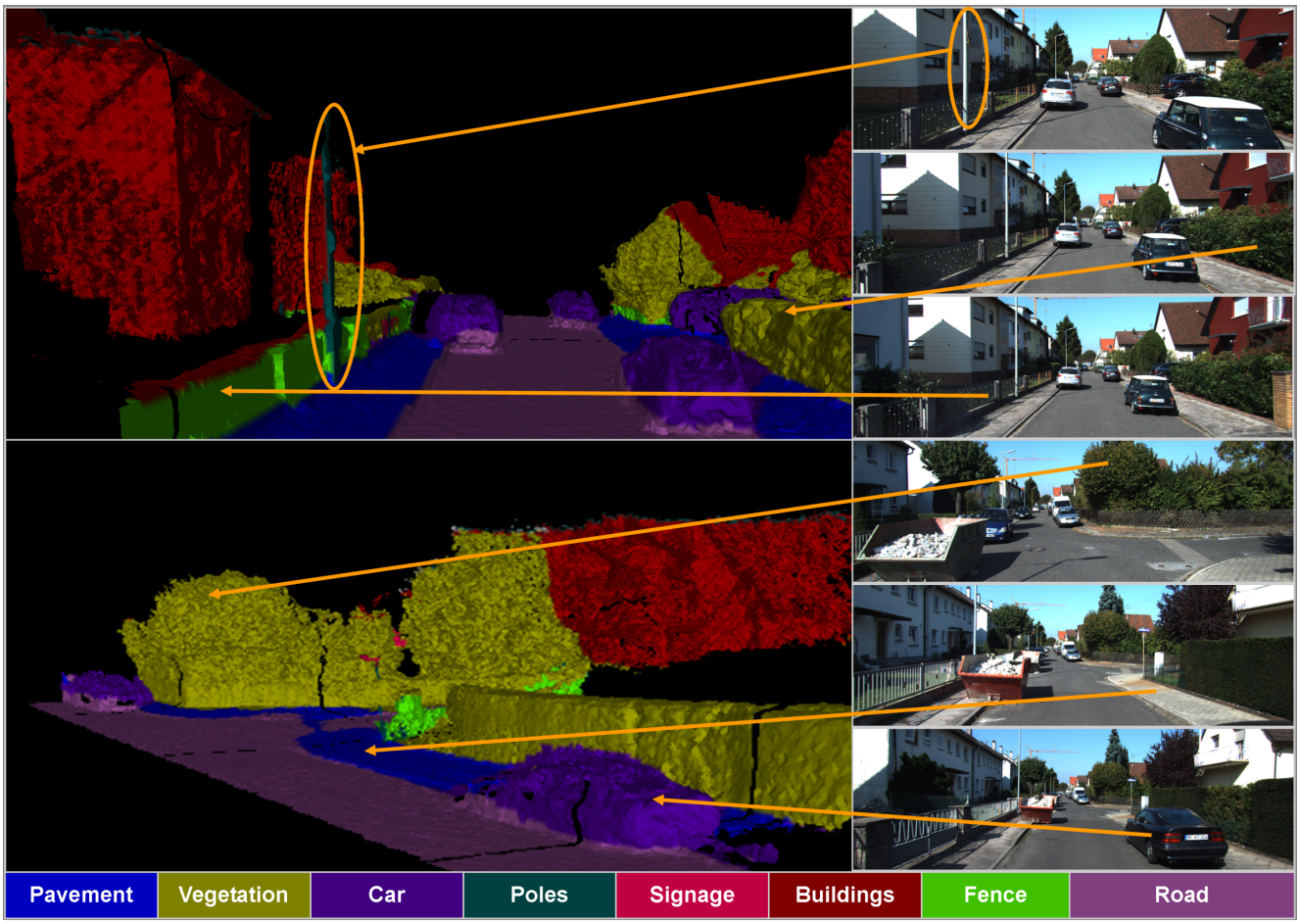

Figure 5. Closeup view of reconstructed semantic model of an urban sequence from KITTI dataset. The arrows relate the position in the model and the associated image.(a) shows the fence, hedge and the post (encircled) in the image and the reconstructed model. (b) shows the right side of the street in the reconstructed model, showing the car, pavement and the trees in for both the cases. (View arrows from bottom, best viewed in colour.)

[20] B. Payne and A. Toga. Surface mapping brain functions on 3d models. In *IEEE Computer Graphics and Applications*, 1990. 2, 3, 5

[21] R. B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *ICRA*, Shanghai, China, May 9-13 2011. 7

[22] S. Sengupta, E. Greveson, A. Shahrokni, and P. Torr. Urban 3d semantic modelling using stereo vision. In *ICRA*, 2013. 2

[23] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context, 2009. ICCV. 1, 2, 3

[24] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor, 2011. ICCV. 1, 2, 4, 5, 6

[25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[26] J. Stuckler, N. Biresev, and S. Behnke. Semantic mapping using object-class segmentation of rgb-d images. In *IROS*, pages 3005–3010, 2012. 1

[27] 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. 2

[28] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection, 2007. PAMI. 3

[29] 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. 1

[30] 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, 2009. 1

[31] H.-P. S. Yutaka Ohtake, Alexander Belyaev. An integrating approach to meshing scattered point data, 2005. ACM Symposium on Solid and Physical Modeling. 1, 2, 3