2015 IEEE Winter Conference on Applications of Computer Vision

A Multi-modal Graphical Model for Scene Analysis

Sarah Taghavi Namin Mohammad Najafi Mathieu Salzmann Lars Petersson
Australian National University (ANU) NICTA[*]
{sarah.namin, mohammad.najafi, mathieu.salzmann, lars.petersson}@nicta.com.au

Abstract

In this paper, we introduce a multi-modal graphical model to address the problems of semantic segmentation using 2D-3D data exhibiting extensive many-to-one correspondences. Existing methods often impose a hard correspondence between the 2D and 3D data, where the 2D and 3D corresponding regions are forced to receive identical labels. This results in performance degradation due to misalignments, 3D-2D projection errors and occlusions. We address this issue by defining a graph over the entire set of data that models soft correspondences between the two modalities. This graph encourages each region in a modality to leverage the information from its corresponding regions in the other modality to better estimate its class label. We evaluate our method on a publicly available dataset and beat the state-of-the-art. Additionally, to demonstrate the ability of our model to support multiple correspondences for objects in 3D and 2D domains, we introduce a new multi-modal dataset, which is composed of panoramic images and LIDAR data, and features a rich set of many-to-one correspondences.

1. Introduction

In this paper, we present a new classification framework which leverages the information of 2D imagery and 3D LIDAR data to perform outdoor scene analysis. The ultimate goal is to obtain a semantic labeling of the entire set of image pixels and 3D points in multi-modal datasets.

In most of the classification frameworks that utilize both 2D and 3D information, appearance and shape features are extracted from corresponding objects and elements (e.g., regions) in the two domains and then inference is performed in only one of the domains [13, 25, 27]. The main shortcoming of these approaches is that their applicability is restricted to the objects and regions that are seen simultaneously in both modalities. In other words, objects which are not captured in one of the modalities are eliminated from the classification process. Furthermore, finding valid and accurate correspondences between the elements of these modalities is a challenging task, due to the difference in the nature of the captured data and the inevitable misalignment between the modalities. This may lead to an association of 2D and 3D features that belong to different objects.

To the best of our knowledge, only two works come closer to treating both modalities separately. In particular, Cadena *et al.* [3] employed a Conditional Random Field (CRF) to perform 2D-3D semantic labeling on 2D RGB images and 3D LIDAR data. They initially designed a graph based on 2D nodes (representing superpixels). Then for each 3D segment, they found its corresponding 2D superpixel and represented them jointly with one single node, as in the aforementioned works. As a result, their design still suffers from similar issues as described above. In contrast, Munoz *et al.* [16] explicitly introduced separate nodes for the 2D and 3D regions and tackled the correspondence problem using an inter-domain overlap function. To be able to handle the uncommon nature of their model, they had to design a specialized co-inference technique, which relies on hierarchical segmentations in both domains and alternates between 2D and 3D labeling. Importantly, they only evaluated their approach on a dataset where one-to-one correspondences between 2D and 3D are available.

In this work, we address the problem of joint 2D-3D outdoor scene analysis by proposing a graphical model in which each region in 2D (superpixel) or 3D (3D segment) is assigned a separate node. The strength of the pairwise link that connect a 2D and a 3D node is adjusted according to the amount of overlap between the 3D segment projected onto the image plane and the 2D image superpixels. The benefits of this representation are threefold. First, it allows us to account for 2D or 3D regions that have no correspondence in the other modality. Second, when a correspondence between a 2D and a 3D region exists, specifying separate nodes addresses the problems that arise because of inaccurate 2D-3D registration and projection. Finally, our representation lets us model

[*] NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the ARC through the ICT Centre of Excellence program.

978-1-4799-6683-7/15 $31.00 © 2015 IEEE
DOI 10.1109/WACV.2015.139

the fact that several superpixels (e.g., from different images) may correspond to a single 3D segment. This yields richer appearance information for the segment and makes inference more reliable. We evaluate our approach on the CMU/VMR urban image + LIDAR dataset [16] and show the superiority of our method over the current state-of-the-art [16].

Furthermore, we release a dataset of panoramic images and 3D point cloud data captured from outdoor scenes (the NICTA/2D3D Dataset[1]). In contrast to the dataset in [16] where all the 3D points have a one-to-one correspondence with 2D image pixels and other points are removed, our data includes the entire set of 3D points which provides naturally occurring many-to-one relationships. Furthermore, for our purpose of multi-modal semantic segmentation, the NICTA/2D3D dataset has the advantage over the KITTI dataset [7] that the point cloud data has both a large vertical and horizontal Field of View (FOV) and is seen from multiple images, thus providing an opportunity to establish correspondences between 3D points and imagery from a large number of view points. This enables research on methods necessary to resolve issues such as correspondence ambiguities, occlusions (either spurious or due to parallax) and missing 2D-3D correspondences. We therefore make use of the NICTA 2D/3D dataset to demonstrate the effectiveness of our method at handling these issues.

2. Related Work

Terrestrial outdoor scene understanding has been investigated using different modalities such as RGB cameras [23] and multi-spectral cameras [18]. In addition to 2D appearance features, 3D information obtained using laser sensors [24, 26] has also received considerable interest in the past few years. Several works have proposed to jointly label and reconstruct the 3D information using videos, multiple view images and stereo systems [6, 12, 21]. However, these types of 3D data are rather noisy and inaccurate for acquiring descriptive features and laser sensors are more reliable devices for capturing 3D information. RGB-D Kinect cameras have also been widely used for indoor scene understanding problems [2, 10, 19] and are shown to perform better compared to the RGB images alone. However, Kinect cameras are not applicable for outdoor scene analysis due to their limited perception range. Considering these issues, one of the most suitable methods of acquiring 2D and 3D information of outdoor scenes is to use conventional camera images along with 3D LIDAR data simultaneously [3, 5, 8, 16].

It has been shown in the past that utilizing the semantic and spatial relationships in 3D or 2D data can significantly enhance the performance of the classification system [1, 9, 14, 22, 25]. This information can be encoded using graphical models. In particular, CRF has often been used to utilize the contextual information of the scene in the form of pairwise and higher order potentials between pixels or superpixels in images, and 3D points (voxels) or segments in 3D data. Pairwise graphical models have been studied in many semantic segmentation problems on 2D or 3D datasets [1, 5, 8, 14, 22, 25]. Due to the limitation of the pairwise models in describing complex contexts in the scene, higher-order graphical models have been used in some works to account for the complicated relationships among the objects and elements [9, 11]. However, as the size of the data and number of nodes in the graph grow, the inference process of higher-order models becomes much more time-consuming.

Only a few methods have proposed to jointly exploit 2D and 3D in graphical models for outdoor scene analysis [3, 5, 8]. In [8] the authors proposed a probabilistic approach for labeling the objects in an urban environment using both laser data and imagery. 3D surface normal features in conjunction with 2D color, texture and geometric features were used to first segment the objects and then classify them into *pavement*, *dirt path*, *smooth wall*, *textured wall*, *vehicle*, *foliage* and *grass*. Douillard *et al.* [5] designed a rule-based system using 3D Velodyne LIDAR data and monocular color imagery to classify the urban environment into 16 different classes.

The major limitation of these works is that the graph is defined over either the 2D domain or the 3D domain and there is no connection between the 2D and 3D nodes. Cadena *et al.* [3] designed a graph where the 2D superpixels and 3D segments which correspond to each other (according to the 3D-2D projection map), were jointly assigned one single node. Then the 2D features were augmented with the 3D features to represent the feature vector of this node in the graph. However, since perfect correspondences between 2D and 3D are assumed, this approach cannot handle many-to-one correspondences, or account for misalignments between the two modalities. Note also that only the 2D results are provided in [3].

Munoz *et al.* [16] assigned separate nodes to 2D superpixels and 3D segments and presented a correspondence function to find the degree of overlap between 2D and 3D nodes and to determine how much they influence each other in the inference process. They also presented a new co-inference technique based on hierarchical segmentations in both the 2D and 3D domains. In their framework, classification was performed in each level of the hierarchy for each domain and the results were transferred to the next hierarchy levels in both domains as a set of fea-

[1] Publicly available at `http://www.nicta.com.au/computer_vision_datasets`.

tures. This approach was repeated through an iterative back-and-forth process over both modalities. Unfortunately, this non-standard model and inference technique make it difficult to generalize the approach to other problems. Furthermore, their method was only demonstrated on a dataset where every 3D point had only one corresponding image pixel.

Our formulation addresses the above-mentioned issues regarding 2D-3D correspondence and is based on a standard inference method, which makes it easy to apply to other similar problems.

3. Method

In this section, we introduce our approach to joint semantic segmentation of 2D panoramic images and 3D point cloud data captured using a LIDAR system. In particular, we consider the scenario where the visual information of an outdoor scene is recorded into F panoramic frames and one 3D point cloud. The ultimate goal is to find the most probable class label for the pixels in the images and 3D points in the point cloud data. In this section, we explain our model which is defined jointly over the 2D and 3D domains.

Given the large size of the point cloud (around $1,000,000$ points) and panoramic images (2000×4000 pixels), it is computationally very demanding to perform inference in a graphical model defined over the entire set of points and pixels. Instead, we build the model based upon image superpixels and 3D segments, as the nodes of the graph.

Here, we propose a full model in which the entire set of 2D superpixels and 3D segments are accounted for in one graph. Fig. 1 illustrates our graphical model. In this figure, squares represent superpixels and spheres represent 3D segments and various types of connections between 2D and 3D nodes are illustrated. Note that some 3D nodes are connected to more than one 2D node, whereas others have no connection with the 2D domain at all.

Let $\mathbf{x}$ be the set of feature computed at the nodes in the graph and $\mathbf{y}$ be the class labels of the nodes. We define the joint distribution over the labels given the features as

$$\mathbf{P}(\mathbf{y^{2D}}, \mathbf{y^{3D}} | \mathbf{x^{2D}}, \mathbf{x^{3D}}) = \frac{1}{Z}\mathbf{exp}\left(-\left(\mathbf{\Phi}^{2D} + \mathbf{\Phi}^{3D} + \mathbf{\Psi}^{2D} + \mathbf{\Psi}^{3D} + \mathbf{\Psi}^{2D-3D}\right)\right), \tag{1}$$

where Z is the partition function. This probability distribution consists of different potentials detailed below.

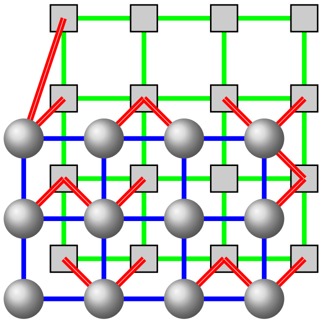


Figure 1. The graphical model in our approach. 2D superpixels are represented by squares and 3D segments are represented by spheres. The blue edges connect 3D segments, green edges link 2D superpixels and double lines (in red) associate the corresponding 2D and 3D nodes. 2D and 3D nodes can be connected to each other, depending on their neighbourhood condition and also the 2D-3D projection.

3.1. 2D Unary Potential

The unary potential for the 2D superpixels is defined as

$$\mathbf{\Phi}^{2D} = \sum_{i=1}^{F} \sum_{j=1}^{N_i} \mathit{\Phi}^{2D}\left(y_{ij}^{2D}, x_{ij}^{2D}\right), \tag{2}$$

where F is the number of image frames and N_i is the number of superpixels in image i. This potential indicates the cost of assigning label y to the j^{th} superpixel in the i^{th} image, given its features, x_{ij}. The $2D$ notation clarifies that this function operates in the 2D domain. The potential function $\mathit{\Phi}^{2D}$ is computed as the negative logarithm of the class probabilities obtained by an SVM classifier.

3.2. 3D Unary Potential

The unary potential for the 3D segments is defined as

$$\mathbf{\Phi}^{3D} = \sum_{j=1}^{M} \mathit{\Phi}^{3D}\left(y_{j}^{3D}, x_{j}^{3D}\right), \tag{3}$$

where M is the number of 3D segments (supervoxels). As for the 2D superpixels, the negative logarithm of the SVM class probabilities are taken as the potential function $\mathit{\Phi}^{3D}$.

3.3. 2D Pairwise Potential

This potential function is defined over all pairwise edges which exist between the adjacent superpixels (j,k). It can be written as

$$\mathbf{\Psi}^{2D} = \sum_{i=1}^{F} \sum_{(j,k)\in E_i^{2D}} \mathit{\Psi}^{2D}\left(y_{ij}^{2D}, y_{ik}^{2D}, x_{ij}^{2D}, x_{ik}^{2D}\right), \tag{4}$$

where E_i^{2D} is the collection of all 2D pairwise edges in frame i, which is generated using the method introduced in [18]. The potential function $\varPsi^{2D}$ is defined in a way that penalizes dissimilar class labels for two adjacent superpixels if their RGB values are very close. It can be expressed as

$$\varPsi^{2D}\Big(\mathrm{y}_1, \mathrm{y}_2, \mathrm{x}_1, \mathrm{x}_2\Big) = \frac{\delta(\mathrm{y}_1 \neq \mathrm{y}_2)}{1 + a\|\mathrm{RGB}_{\mathrm{x}_1} - \mathrm{RGB}_{\mathrm{x}_2}\|_1} . \tag{5}$$

This potential equals zero (via the delta indicator function) if the pair of superpixels have identical class labels. a is the weight of the RGB contrast which is determined using cross-validation ($a = 0.05$).

3.4. 3D Pairwise Potential

This potential function is defined as

$$\mathbf{\Psi}^{3D} = \sum_{(j,k)\in E^{3D}} \varPsi^{3D}\Big(y_j^{3D}, y_k^{3D}, x_j^{3D}, x_k^{3D}\Big) \tag{6}$$

Here, E^{3D} denotes the collection of pairwise edges between 3D segments. We consider every pair of segments whose minimum inter-point distance is lower than a threshold ($D_t = 1\mathrm{m}$) to be a neighbor and constitute a pairwise connection. The potential function $\varPsi^{3D}$ is computed as

$$\varPsi^{3D}\Big(\mathrm{y}_1, \mathrm{y}_2, \mathrm{x}_1, \mathrm{x}_2\Big) = \frac{\delta(\mathrm{y}_1 \neq \mathrm{y}_2)}{1 + b|\theta_{\mathrm{x}_1} - \theta_{\mathrm{x}_2}|} , \tag{7}$$

where θ is the angle between the direction of the average normal vector of 3D segment and the vertical axis. The weight b is optimized by cross-validation ($b = 1/90$).

3.5. 2D-3D Pairwise Potential

This potential is applied to the edges that connect 2D nodes to 3D nodes. It can be written as

$$\mathbf{\Psi}^{2D-3D} = \sum_{i=1:F,(ij\ k),\in E^{2D-3D}} \varPsi^{2D-3D}\Big(y_{ij}^{2D}, y_k^{3D}, x_{ij}^{2D}, x_k^{3D}\Big) . \tag{8}$$

Since the outdoor scene is captured using a panoramic RGB camera and a 360° laser scanning system, many objects and regions can be observed in both data modalities. The pairwise potential $\mathbf{\Psi}^{2D-3D}$ takes the relationships between the 2D objects and their 3D counterparts into account by considering pairwise links between them. To find all the pairwise links between the 2D and 3D domains, the point cloud is projected onto the image planes. As a result, the projection of each 3D segment may intersect with zero, one or more superpixels in the images. Some segments may also be observed in more than one panoramic image. The list of the entire set of 2D-3D pairwise links is recorded into E^{2D-3D} and the potential function $\varPsi^{2D-3D}$, which is computed as

$$\varPsi^{2D-3D} = w_{ij,k}\delta(\mathrm{y}_1 \neq \mathrm{y}_2) , \tag{9}$$

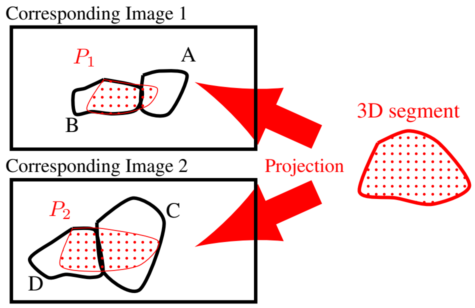


Figure 2. Illustration of how the corresponding superpixels of each 3D segment and their 2D-3D pairwise weights are determined. A 3D segment is projected onto its nearby image planes. A pairwise link is created for each superpixel that has a significant overlap with this projection. The weight of this link is determined according to the amount of overlap with the projected 3D segment and the size of the superpixel.

is applied to all of them. In this function, $w_{ij,k}$ is the 2D-3D overlap weight and is calculated as follows. The size of the overlap between the projected 3D segment and each of its corresponding superpixels is computed and normalized with respect to the size of the superpixels. This yields an overlap weight vector $\mathbf{w} = [w_1, w_2, \dots, w_t]$ for each 3D segment, where t is the number of overlapping superpixels.

Fig. 2 illustrates how this overlap weight vector is computed where a 3D segment in the point cloud data has some projection overlap with the superpixels in two different images. In this figure, suppose that we have $\frac{P_1 \cap A}{A} = 0.2$, $\frac{P_1 \cap B}{B} = 0.8$, $\frac{P_2 \cap C}{C} = 0.6$ and $\frac{P_2 \cap D}{D} = 0.6$. This yields the overlap weight vector $\mathbf{w} = [0.2, 0.8, 0.6, 0.6]$. As a result, the superpixel with maximum size of overlap impacts the 3D segment more than others (due to a larger weight for its 2D-3D pairwise link).

Note that due to the imperfection of 2D segmentation, there might be some cases where the projection of thin or small objects (like *poles* or *wires*) is surrounded by only one large superpixel (Fig. 3). As a consequence, the normalized overlap weight for these cases becomes very small which makes the impact of the 2D-3D pairwise potential for them negligible. To overcome this issue, the overlap weights are normalized again (by dividing them by the maximum weight among all corresponding superpixels of each 3D segment). For instance, the weight vector $\mathbf{w}$ which was computed for Fig. 2 is normalized by dividing all its components by their maximum value of 0.8 which results in $\mathbf{w} = [0.25, 1, 0.75, 0.75]$.

4. Data

In this section, we first review two existing multi-modal 2D/3D datasets and discuss their problems as benchmarks for semantic segmentation. Then we present a new multi-modal dataset in which those problems have

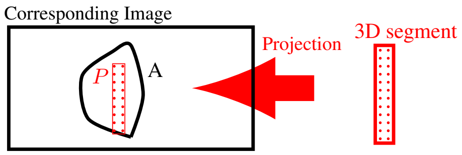


Figure 3. Second normalization of the 2D-3D pairwise weight vector. Since the object is very thin, the ratio $\frac{P \cap A}{A}$ weakens the pairwise link between the 3D segment and its only counterpart superpixel in the image. Therefore, we normalize this ratio w.r.t. the size of the overlap.

been addressed.

KITTI [7] is probably the largest publicly available multi-modal dataset which is mainly used for object detection tasks. The main problem with KITTI is the small vertical FOV of the point cloud data. As shown in [3], a large portion of the images in this dataset does not correspond to any laser scanning data. As a consequence, 2D labeling relies mostly on image data rather than on multi-modal 2D/3D data.

CMU/VMR is another multi-modal dataset which is composed of wide RGB images in conjunction with 3D point cloud data, collected from 372 urban scenes [16]. One issue with this dataset is that the 3D data is not complete and the points with no corresponding image pixels have been removed. The 3D points were annotated by back-projecting the labeling of the annotated images. However, due to the inaccurate 2D-3D back-projection, the ground truth of some of the 3D points is incorrect.

In this paper, we present a new multi-modal dataset (NICTA/2D3D dataset) for outdoor scene understanding which consists of a series of 2D panoramic images with corresponding 3D LIDAR point clouds. It contains 12 outdoor scenes, each of which includes a block of 3D point cloud together with several panoramic images. The number of 3D points in the scenes varies from 1 to 2 millions, and each scene contains between 10 and 20 panoramic images depending on the size of the point cloud block.

The dataset was manually annotated in the 3D domain and the ground truth labeling of the panoramic images were obtained via 3D-2D projection of the 3D labels. The 2D ground truth images were later checked and retouched to produce a more precise 2D ground truth. This step accounts for projection errors due to misalignments or parallax and also deals with moving and/or reflective objects whose point cloud data is very sparse. Additionally, the label *Sky*, which does not exist in the 3D data, was included as a new label in the 2D images. The point cloud data is seen from multiple viewpoints thanks to the 360° panoramic images, and its FOV covers the entire image (both vertically and horizontally) instead of just a portion in the KITTI dataset [7]. In contrast to the dataset in [16], where every 3D point corresponds to a 2D image pixel (i.e., 3D points not satisfying this property were removed), the NICTA/2D3D data includes the entire set of 3D points.

The second advantage of the NICTA/2D3D data over the aforementioned datasets is that the panoramic images provide an opportunity to capture each object several times in different frames and from different viewpoints. Therefore it not only provides the corresponding 2D information for each 3D segment, but also does it several times from different views. We benefited from this property in our graph and connected each 3D node to all of its corresponding 2D nodes in different frames. As a result, each 3D segment can leverage the appearance features of the scenes from several views and each 2D region can utilize the 3D information as well as the 2D information of other superpixels (in other panoramic image frames), which are indirectly connected to it in the graph via its corresponding 3D node.

5. Experiments

In this section, we first describe the steps required to compute the unary potentials (3D and 2D), and then discuss our experiments on two datasets in details.

5.1. 3D Features and Unary Potentials

We employed Fast Point Feature Histogram (FPFH) [20], eigenvalue descriptors, vertical axis deviation and height as 3D shape features, extracted from the point cloud data. The FPFH descriptors [20] encode the relationships between the 3D points and their neighbors in terms of the spatial distance and orientation of surface normal vectors. The eigenvalue descriptors are a good measure of linearity and planarity of the surface on which the 3D point lies. The vertical axis deviation helps discriminating vertical planar surfaces from horizontal surfaces. Finally, the height of each point can provide a clue about the object.

A probabilistic SVM classifier was trained and applied to the extracted features. The 3D points were classified into one of the L pre-defined class labels and the probability of belonging to each class was also computed for each point. We then separated the points into L different groups according to their class labels and performed a k-means segmentation to each group to further divided them into our final 3D segments. The class probabilities of each 3D segment was calculated by averaging over the class probabilities of its points. Finally, the unary potentials of the 3D segments were computed by taking the negative logarithm of their class probabilities, as discussed in Section 3.2.

5.2. 2D Features and Unary Potentials

In contrast to the previous step, where classification was performed at the level of individual 3D points, the

2D classifier is directly applied to the superpixels. We utilized the MeanShift method of [4] to segment the images into superpixels. We then extracted a histogram of SIFT features [15], average RGB values and GLCM features (Energy, Homogeneity and Contrast) from each superpixel. The visual words for the SIFT histogram were obtained from half of the training images. These features were used to train a probabilistic SVM classifier and predict the class probabilities of the superpixels in the test images. The 2D unary potentials for each superpixel was then computed as the negative logarithm of the attained class probabilities (see Section 3.1).

5.3. Experimental Results on CMU/VMR

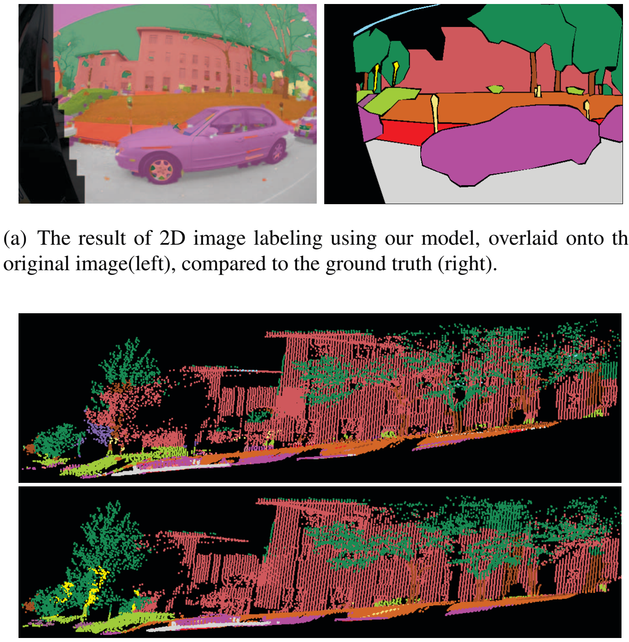
As mentioned in Section 4, the CMU/VMR dataset consists of 372 urban scenes, each of which was surveyed using an RGB camera and a laser scanner. We divided the dataset into 4 non-overlapping folds, one of them used for validation and the others used for 3-fold cross-validation. The CRF parameters were obtained from the validation data, and Loopy Belief Propagation [17] was chosen as the inference method. We use the F1-score as the performance measure in order to facilitate the comparisons with [16]. On average, we achieve $F1^{2D} = 0.46$ for the semantic segmentation of the images and $F1^{3D} = 0.43$ for the semantic segmentation of the 3D point cloud, which are higher than the results in [16] who reports $F1^{2D} = 0.45$ and $F1^{3D} = 0.39$. Table 1 and Table 2 show our quantitative 2D and 3D classification results per class, compared to the results in [16].

Since the ground truth of the 3D data in this dataset is not as reliable as the 2D ground truth (as described in Section 4), and also because the 3D point cloud is not as dense as it could be (due to the removal of the 3D points with no pixel correspondence), the results of the 3D semantic labeling is lower than that of 2D semantic segmentation. This issue is even more critical for the last five categories in Table 2, which are all thin and small objects and thus influenced the most by projection errors. Nevertheless, as indicated in Table 2, our model has considerably improved the performance of the 3D labeling for these classes. The average F1-score of the 3D labels has also been improved from 25% to 43%, which demonstrates the benefits of our model.

Fig. 4 depicts the qualitative results for a sample scene in this dataset, compared to the ground truth.

5.4. Experimental Results on NICTA/2D3D

For out new dataset, we used the same experimental setup (number of folds, parameter selection, inference method, performance measure) as in the previous experiment on the CMU/VMR dataset. The average F1-scores of our model are $F1^{2D} = 0.45$ and $F1^{3D} = 0.52$, which



(a) The result of 2D image labeling using our model, overlaid onto the original image(left), compared to the ground truth (right).

(b) The result of the 3D point cloud semantic labeling (above), compared to the ground truth (below).

Figure 4. The qualitative results of our proposed model for semantic segmentation of (a) 2D image and (b) 3D point cloud, captured from a scene in CMU/VMR dataset. The color codes for this figure are: White=Road, Brown=TreeTrunk, Light-Red=Building, Green=TreeTop, Light-Green=Shrub, Pink=Vehicle, Red=Sidewalk, Orange=Ground, Yellow=Utility pole

outperform the results of the unary classifier ($F1^{2D} = 0.38$ and $F1^{3D} = 0.43$).

Table 3 and Table 4 evidence the fact that the performance of the system in each domain has been improved by incorporating information from the other domain via 2D-3D pairwise links. In particular, our model has increased the classification rate of the classes which had a small number of training samples, by exploiting the 2D-3D multi-correspondence. Note that 2D-3D pairwise edges can add inter-domain semantic information while they do not yield over-smoothing. As evidenced by Table 3, the 2D pairwise edges were unable to recover any misclassified objects in the *Post* and *Barrier* categories. Nonetheless, our model has improved the classification rate for these classes by 8% and 5%, respectively.

Table 3 and Table 4 show that except for two classes (*Road* and *Sidewalk*), the other classes have had a significant improvement in their F1-score in at least one of the 2D or 3D datasets. This is mainly because the 3D information of *Road* and *Sidewalk* are very similar, and the 2D-3D pairwise edges are not effective enough to correct the misclassification between them in the 3D data. As can be seen in Table 3, this misclassification has been trans-

Table 1. The F1-scores of the 2D classification for the CMU/VMR dataset ([16]) using our model compared to the method in [16].

	Road	**Sidewalk**	**Ground**	**Building**	**Barrier**	**Bus stop**	**Stairs**	**Shrub**	**Tree trunk**	**Tree top**	**Small Vehicle**	**Big vehicle**	**Person**	**Tall light**	**Post**	**Sign**	**Utility pole**	**Wire**	**Traffic Signal**	**avg**
Unary	95	81	75	56	29	17	32	50	31	53	32	49	29	16	15	**16**	33	41	29	41
Pairwise 2D	95	82	76	64	28	17	24	52	**34**	73	51	**54**	28	**17**	**16**	13	33	**46**	28	44
Our Model	94	87	**79**	74	45	**22**	**40**	54	27	84	67	24	38	13	2	10	**37**	35	**40**	**46**
Munoz [16]	**96**	**90**	70	**83**	**50**	16	33	**62**	30	**86**	**84**	50	**47**	2	9	**16**	14	2	17	45

ferred to the 2D domain via 2D-3D pairwise edges and has deteriorated the F1-scores of *Sidewalk* and *Road* in the 2D dataset.

The qualitative results of the 2D and 3D semantic labeling in two different scenes of the NICTA/2D3D dataset are illustrated in Fig. 5.

6. Conclusion

In this paper, we have proposed a graphical model that enabled us to perform a joint inference on two different modalities of data (2D imagery and 3D point cloud) and improve the semantic labeling in both modalities. We have incorporated 2D-3D pairwise edges in the graph (in addition to 2D-2D and 3D-3D edges) that connect corresponding 2D nodes and 3D nodes and transfer information from one modality to the other. Although these pairwise connections utilize information from both the 2D and 3D domains to enhance the labeling of corresponding superpixels and 3D segments, they do not force these corresponding nodes to be assigned identical class labels, which is beneficial in the presence of projection errors. As a result, such pairwise connections do not cause over-smoothing and improve the performance of the system, especially for small and narrow classes. Our experiments have evidenced that we outperform the state-of-the-art on a publicly available dataset. Furthermore, we have introduced a new publicly available multi-modal dataset, which addresses the problems of the existing datasets in terms of correspondence between the 2D and 3D domains. Our model can be applied to data with other modalities, given that connections between the data modalities are well-defined.

References

[1] A. Anand, H. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for 3d point clouds. *IJRR*, 2012.

[2] L. Bo, K. Lai, X. Ren, and D. Fox. Object recognition with hierarchical kernel descriptors. In *CVPR*, pages 1729–1736. IEEE Computer Society, 2011.

[3] C. Cadena and J. Koseck. Semantic Segmentation with Heterogeneous Sensor Coverages. In *ICRA*, 2014.

[4] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603–619, May 2002.

[5] B. Douillard, A. Brooks, and F. Ramos. A 3D laser and vision based classifier. In *ISSNIPC*, 2009.

[6] G. Floros. Joint 2d-3d temporally consistent semantic segmentation of street scenes. In *CVPR*, pages 2823–2830, 2012.

[7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.

[8] P. N. Ingmar Posner, Mark Cummins. Fast probabilistic labeling of city maps. In *RSS*, Zurich, Switzerland, June 2008.

[9] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009.

[10] H. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3D point clouds for indoor scenes. In *NIPS*, pages 244–252, 2011.

[11] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Inference methods for crfs with co-occurrence statistics. *IJCV*, 103(2):213–225, 2013.

[12] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. F. Clocksin, and P. H. S. Torr. Joint optimization for object class segmentation and dense stereo reconstruction,. *IJCV*, 100(2):122–133, 2012.

[13] K. Lai, L. Bo, X. Ren, and D. Fox. Detection-based object labeling in 3d scenes. In *ICRA*, 2012.

[14] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, December 2013.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, Nov. 2004.

[16] D. Munoz, J. A. Bagnell, and M. Hebert. Co-inference for multimodal scene analysis. In *ECCV*, pages 668–681, 2012.

[17] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, pages 467–475, 1999.

[18] M. Najafi, S. Namin, and L. Petersson. Classification of natural scene multi spectral images using a new enhanced crf. In *IROS*, pages 3704–3711, Nov 2013.

[19] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[20] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms for 3d registration. In *ICRA*, pages 1848–1853, 2009.

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

[22] R. Shapovalov, A. Velizhev, and O. Barinova. Non-associative markov networks for 3d point cloud classification. In *PCVIA*. ISPRS, 2010.

[23] I. Tang and T. P. Breckon. Automatic Road Environment Classification. *ITITS*, 12:476–484, 2011.

[24] A. Teichman, J. Levinson, and S. Thrun. Towards 3d object recognition via classification of arbitrary object tracks. In *ICRA*, pages 4034–4041. IEEE, 2011.

[25] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *ICCV*, pages 686–693. IEEE, 2009.

[26] X. Xiong, D. Munoz, J. A. Bagnell, and M. Hebert. 3-d scene analysis via sequenced predictions over points and regions. In *ICRA*, 2011.

[27] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. In *ECCV*, volume 6314, pages 708–721, 2010.

Table 2. The F1-scores of the 3D classification for the CMU/VMR dataset ([16]) using our model compared to the method in [16].

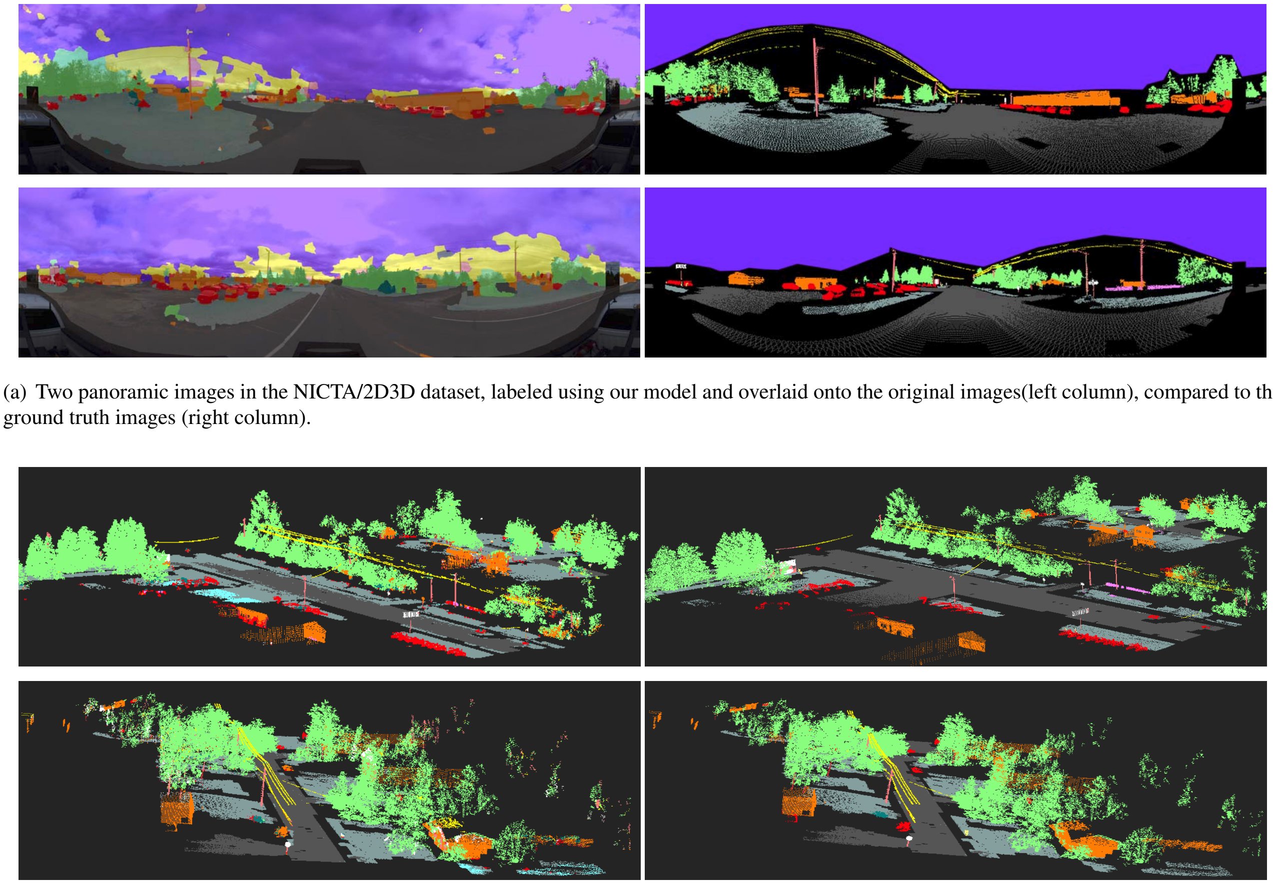
	Road	Sidewalk	Ground	Building	Barrier	Bus stop	Stairs	Shrub	Tree trunk	Tree top	Small Vehicle	Big vehicle	Person	Tall light	Post	Sign	Utility pole	Wire	Traffic Signal	avg
Unary	70	49	62	67	34	2	19	26	11	67	34	4	13	2	0	1	2	0	0	24
Pairwise 3D	71	51	64	66	37	2	22	25	11	67	33	4	13	2	0	1	2	0	0	25
Our Model	**92**	**85**	**81**	85	**50**	**16**	**42**	55	**29**	82	70	16	**43**	6	**2**	**7**	**29**	**9**	**23**	**43**
Munoz [16]	82	73	68	**87**	46	11	38	**63**	28	**88**	**73**	**56**	26	**10**	0	0	0	0	0	39

Table 3. The F1-scores of the 2D classification for the NICTA/2D3D dataset using our model.

	Grass	Building	Tree trunk	Tree leaves	Vehicle	Road	Bush	Pole	Sign	Post	Barrier	Wire	Sidewalk	Sky	avg
Unary	**80**	33	14	80	49	**95**	16	28	3	0	0	29	15	**98**	38
Pairwise 2D	69	37	17	78	55	90	14	29	6	0	0	**32**	**47**	97	41
Our Model	74	**56**	**21**	**82**	**58**	92	**23**	**33**	**19**	**8**	**5**	**32**	29	97	**45**

Table 4. The F1-scores of the 3D classification for the NICTA/2D3D dataset using our model.

	Grass	Building	Tree trunk	Tree leaves	Vehicle	Road	Bush	Pole	Sign	Post	Barrier	Wire	Sidewalk	Sky	avg
Unary	52	61	27	87	58	**82**	10	24	19	43	19	74	0	#	43
Pairwise 3D	55	**81**	34	95	61	62	19	**40**	**29**	42	22	86	0	#	48
Our Model	**63**	**81**	**41**	**96**	**70**	76	**21**	38	28	**47**	**23**	**87**	0	#	**52**



(a) Two panoramic images in the NICTA/2D3D dataset, labeled using our model and overlaid onto the original images(left column), compared to the ground truth images (right column).

(b) Two point cloud blocks in the NICTA/2D3D dataset, labeled using our model (left column), compared to their ground truth (right column).

Figure 5. The qualitative results of our proposed model for semantic segmentation of (a) 2D images and (b) 3D point cloud, captured from two different scenes in the NICTA/2D3D dataset. The color codes for this figure are: Dark-Gray=Road, Orange=Building, Green=Leaves, Red=Vehicle, Blue=Sidewalk, Gray=Grass, Light Pink=Pole, Purple=Sky, Yellow=Wire.