CRF based method for Curb Detection using semantic cues and stereo depth

Danish Sodhi*
RRC, KCIS,
IIIT Hyderabad, India

Sarthak Upadhyay†
RRC, KCIS,
IIIT Hyderabad, India

Dhaivat Bhatt‡
RRC, KCIS,
IIIT Hyderabad, India

K Madhava Krishna
RRC, KCIS,
IIIT Hyderabad, India

Shanti Swarup
Uurmi Systems

ABSTRACT

Curb detection is a critical component of driver assistance and autonomous driving systems. In this paper, we present a discriminative approach to the problem of curb detection under diverse road conditions. We define curbs as the intersection of drivable and non-drivable area which are classified using dense Conditional random fields(CRF). In our method, we fuse output of a neural network used for pixel-wise semantic segmentation with depth and color information from stereo cameras. CRF fuses the output of a deep model and height information available in stereo data and provides improved segmentation. Further we introduce temporal smoothness using a weighted average of Segnet output and output from a probabilistic voxel grid as our unary potential. Finally, we show improvements over the current state of the art neural networks. Our proposed method shows accurate results over large range of variations in curb curvature and appearance, without the need of retraining the model for the specific dataset.

CCS Concepts

•**Computing methodologies** → **Computer vision;**

Keywords

Stereovision, Conditional Random Field, Deep learning, Curbs

1. INTRODUCTION

Curb detection is a critical component for autonomous driving and driver assistance systems. Safe navigation of the vehicle is the most crucial component in autonomous driving. Majority of road accidents and fatalities are the due to erroneous human driving. Autonomous driving can help in reducing the accidents due to human errors. Robust detection of curb boundaries will enable any autonomous system to determine navigable space in the scene.

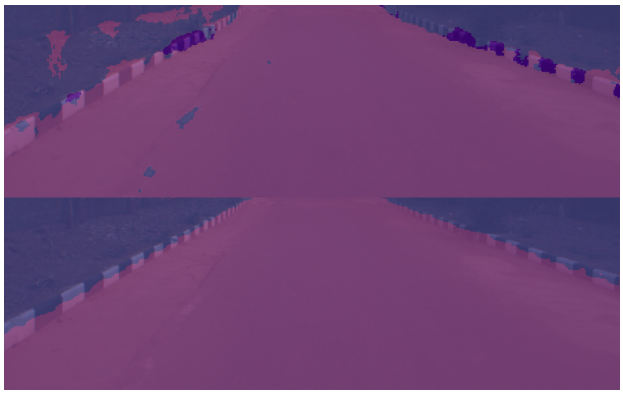

Figure 1: Segmentation of drivable area using SegNet(top) and proposed method(bottom)

Curbs usually exhibit various heights, have different appearances and often varying curvatures that makes the task of curb detection more challenging than lane markings detection. In this paper, we propose a robust curb detection approach that will help in estimating drivable area for intelligent vehicles. In our approach, we use an existing pre-trained deep neural network for semantic segmentation of the road-scene. Indian roads show diverse variations in road appearance. With dirt, mud and illumination variations, it is a tedious task to train a neural network every time autonomous vehicle goes to a different environment. We use the SEGNET architecture [1] to obtain per pixel semantic labelling of the scene, which provides for one of the two unary potential terms. A temporary consistency term across images enabled by Visual Odometry [4] forms the other unary potential. The pairwise potentials come from height, shape and appearance features. A dense CRF is formulated and the mean field inference provides for a highly accurate and long range (40 meters) segmentation of drivable and non

*first author
†second author
‡second author

ICVGIP, December 18-22, 2016, Guwahati, India

DOI: http://dx.doi.org/10.1145/3009977.3010058

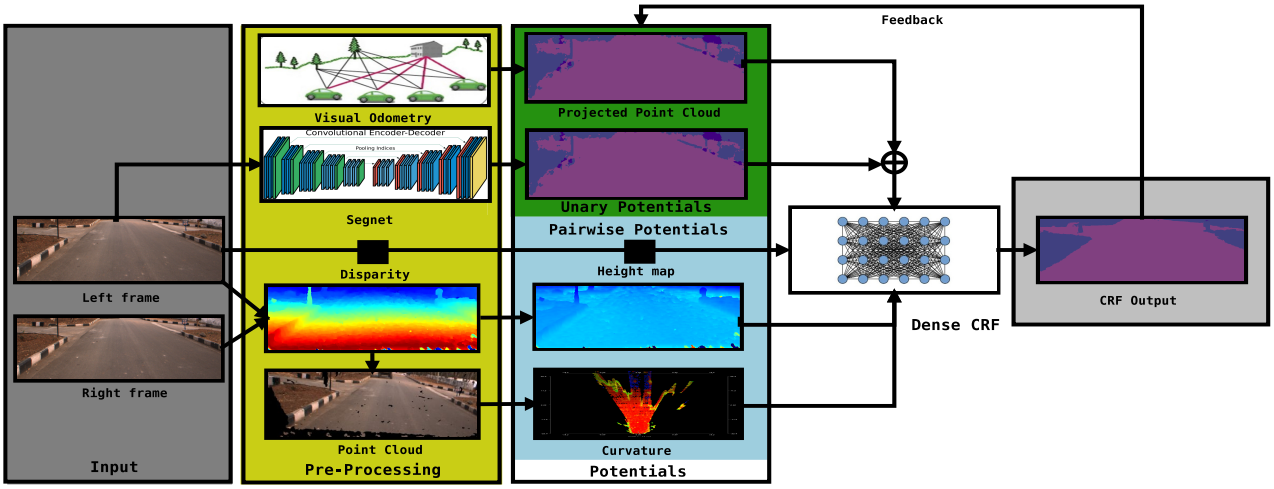

Figure 2: This is a graphical representation of proposed pipeline. Stereo frames are used to generate disparity map. Disparity map was computed with respect to the left frame. Left frame has been used to generate pixel-level semantic segmentation. Height profile and Curvature are estimated using a point cloud. Height map, curvature are used as features to compute pairwise potentials of dense CRF. Similarly, Segnet output and visual odometry are used to estimate unary potentials. The pipeline generates a more accurate semantic segmentation for drivable and non-drivable area.

drivable portions of the road.

Our main contributions are as follows:

- We effectively combine semantic, shape, height and temporal consistency cues through a dense CRF formulation for long range curb detection
- A probabilistic framework for the task of detection of drivable area boundaries.
- Combination of both appearance and 3D reasoning for a more accurate segmentation even at a large distance.
- Tackle the sparsity of the point cloud by employing novel techniques over the acquired features.
- The high fidelity segmentation consequent of the CRF formulation precludes the need to fine tune a pre-trained network classifier for every new set of road scenes an autonomous vehicle or car would encounter. This claim is vindicated as an improvement over baseline CNN classifiers such as SEGNET [1] are presented in the results section. Specifically despite the pre-trained network getting confused on Indian scenes (as pavement labels in European conditions get confused with road labels in Indian scenes) we are able to recover from such erroneous semantic labelling.
- With respect to most recent curb detection methods [2] more than 30% increased curb lengths are segmented by the proposed method.

The paper is divided in 3 parts, in section 2 we talk about the related work on curb detection. In section 3 we present out approach along with its various components. Finally we present our results in section 4, and final conclusion.

2. REALTED WORK

There are several approaches that address the problem of curb detection using different sensors.[13] proposes a vision

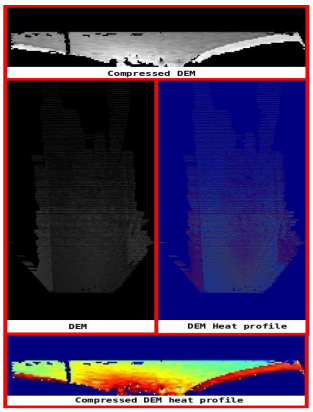

Figure 3: Result of [8] [11] failing for sparse stereo data.

based approach for curb detection using cluster of parallel lines. This approach is solely based on intensity variation and could be misleading as straight lines in an image might be a potential candidate for lane markings rather than curbs. [9] presents a curb detection approach using 3D data from dense stereo. 3D data points are mapped into a digital elevation map in bird's eye view. Canny edge detection followed by Hough transformation is applied on the generated elevation map to get the potential candidate for the curbs. However this approach mainly detects straight line curbs and fails to detect curved curbs. This approach is limited for a small distance only. [7] also extracts potential curb candidates using edge detection on DEM and fits polynomial using RANSAC. However, these approaches fail due to sparsity of the stereo data. The problem of sparsity in point cloud is partially taken care in [8] where height is propagated along the longitudinal direction, but this method leads to generation of false curb candidates for curved roads. There is another proposal [11] which deals with generation of compressed version of DEM. This method divides cam-

era's horizontal field of view into polar slices with constant aperture (figure 3). But expansion of this compressed space leads to erroneous curb boundaries.[10] further improves the ideas proposed in [9] by temporal filtering of curb points to remove the false positives. In [2], a curb detection approach based on curvature features and CRF is proposed. Nearest neighbouring points are used to create a weighted covariance matrix corresponding to each pixel. Eigenvalues of these matrices are used to compute curvature value at that pixel.[14] presents an approach for curb detection based on conditional random fields (CRF). 3D points from dense stereo are assigned to adjacent surfaces of curbs, mainly street and sidewalk using Conditional random fields.

Most of the current methods to detect curbs are solely based on 3D data. Our method proposes a graph based minimization approach that combines depth and appearance information. Additional clue of appearance yields accurate segmentation for longer distances where point cloud data become less reliable due to sparsity.

3. METHODOLOGY

This section briefly walks you through the pipeline of our proposed method (figure 2). Our method consists of following steps:

The first step of the pipeline is to generate semantic segmentation of a raw image using deep neural network. There are two semantic labels for each image frame corresponding to drivable and non-drivable area. The second step involves computing disparity using semi-global block matching [5]. Disparity is further used to generate height elevation map in camera's frame of reference. Semantic labels and height information computed in previous steps are then used to estimate potentials for the cost function of Conditional random field. CRF results in refining the segmentation results obtained from Segnet. Further, we apply a temporal filter has been integrated in the pipeline to improve accuracy of the prediction. Introducing temporal filter discards false positives occurring near ambiguous regions.

3.1 Stereo Depth and pre processing

We define the world frame (X_i, Y_i, Z_i) centred at the optical center of left camera .The XZ plane is parallel to the road plane in such a way that X-axis points towards right ,Z-axis points towards front and Y-axis points in downwards direction as shown in figure 4 . Given left and right image corresponding to each frame , we use robust semi-global block matching cite to calculate a disparity map where each pixel (u_i, v_i) of the map represents the disparity value for that pixel in left image. Using the disparity map , each cell (u_i, v_i) in the image is then assigned an elevation value corresponding height Y_i at that pixel. Height corresponding to each pixel is calculated using the projection matrix as

$$Z_i = f \cdot B/d \tag{1}$$

$$Y_i = \frac{(v_i - c_y) \cdot Z}{f_y} \tag{2}$$

where f_y, c_y are the camera focus and center in y-axis respectively. B is defined as the baseline of the stereo camera pair with a focal length of f.

3.2 Segnet on images and label merging

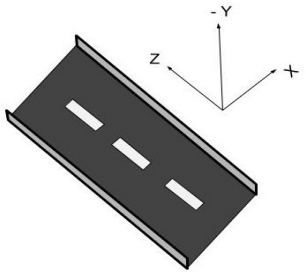

Figure 4: Camera Co-ordinate System

Current advancements in deep learning have enabled research community to utilize pretrained model over a large variety of applications. With neural networks achieving state-of-the art performance, it has been widely used for number of computer vision applications. In our method, we use various deep neural networks for pixel-wise semantic segmentation. SegNet is one such segmentation engine. Segnet is a deep convolutional encoder-decoder architecture primarily inspired to solve road-scene understanding problems. SegNet has three essential components, encoder network, corresponding decoder network and a classification layer for assigning label to the pixel among a set of several classes.

Encoder: Encoder network of SegNet has 13 convolutional layers. The architecture is similar to the architecture of VGG16 [15] network designed to classify object. Each encoding layer in the encoder network is responsible for generating a set of feature maps by performing convolution with a filter bank. Subsequently, pooling and subsampling is performed to achieve translation invariance and global context for each pixel in the feature map.

Decoder: Each encoder layer has its corresponding decoding layer. Hence, decoder network has 13 decoding layers. The decoder maps low resolution encoder feature output to dense feature maps. Internally, it uses pooling indices computed in the corresponding encoder layer to perform upsampling, which are convolved with trainable filters to produce dense feature maps.

Classification layer: The high dimensional dense features coming from final layer of the decoding network is fed to a multi-class soft-max classifier. The output of the soft-max classifier is class probabilities estimated for each pixel independently. The class with highest probability value gets assigned as a predicted class for each pixel.

In this paper, we show an improvement over the output of the deep neural networks by fusing it with the 3D depth information from the stereo cameras. The 3D and colour information is used to correct the semantic labelling given by the output of the deep neural model, hence eliminating the need to train a network primarily for this task.

3.3 Curvature estimation from pointcloud

The proposed curb detection method is based on surface curvature estimation presented in [12]. This feature has been also used in [2] and [3] for free space detection. The curvature describes the variation along the surface normal and it varies between 0 and 1, where low values correspond to flat surfaces. The curvature feature is more robust and stable

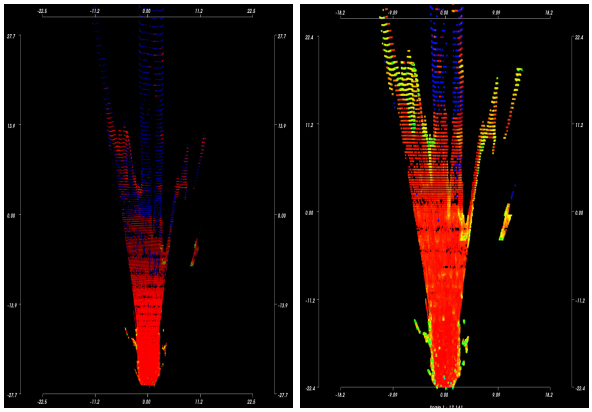

Figure 5: Curvature results using method described in Section 3.3 compared to [3], here red are low curvature points, yellow are points with high curvature and blue are points with very low neighbours.

than tangent plane normal vectors. For each point p, the nearest neighbours (NN) p_i in a surrounding area defined by a radius R are selected. These points are used to create a covariance matrix, where k denotes the number of nearest neighbours.

$$\overline{p} = \frac{1}{k}\sum_{1}^{k} p_i \tag{3}$$

$$C = \sum_{1}^{k}(p_i - \overline{p}) \cdot (p_i - \overline{p})^T \tag{4}$$

The eigenvector V and eigenvalues λ of C are computed as $C \cdot V = \lambda \cdot V$. A curvature measure γ is defined by equation [5], where $\lambda_0 \leq \lambda_1 \leq \lambda_2$ are the eigenvalues of the covariance matrix C. Instead of λ we use σ which is defined as $\sigma = \sqrt{\lambda}$. Finally curvature is defined as

$$\gamma = \frac{\sigma_0}{\sigma_0 + \sigma_1 + \sigma_2} \tag{5}$$

We found this to be more discriminative than the the features used in [3]. We also use a dynamic radius of search described in section 3.5.

3.4 Visual Odometry

Visual odometry is the process of determining the position and orientation of a robot by analyzing the associated camera images. We use method proposed by [4] to compute the relative translation(T) and rotation(R) between images. The computed transformation matrix is used to project the 3D points from previous frame to the current frame as follows

$$P_i^t = [RT] \cdot P_i^{t-1} \tag{6}$$

here P_t^i is the point in the point cloud.

This is done enforce a temporal consistency by maintaining a probability P_i with each point on the image. P_i here is the probability that the point corresponds to the drivable area. The probabilistic update is done as below

$$P_i = P(l_{t-1}) \cdot P(l_t|l_{t-1}) \tag{7}$$

The mapping is performed using a visual slam on the input images. The points are projected to the new camera frame as given below.

$$x = P \cdot X \tag{8}$$

here X is the 3D point in the camera frame and P is the camera projection matrix. A bilateral filter is applied on the resultant image to compute the missing values due to occlusion or other reasons.

3.5 Sparsity of Point Clouds

The major issue with stereo and Lidar point clouds is, as with distance increases the density of point cloud decreases. [3] handles this by merging multiple scans using Iterative Closest point algorithm. The problem with this approach is that at large distances the noise in a single scan is high and gets accumulate as we merge more scans, thereby corrupting the curvature and height observations. We overcome this by introducing two interesting techniques.

Firstly, the radius of search in curvature computation in section 3.3 is formulated as a linear function of depth.

$$R(p_i) = R_c + C \cdot p_i(Z)$$

where p_i is the point in the point cloud and $p_i(Z)$ is the corresponding depth. R_c is the constant starting radius radius set to 0.25m. We get best results for C = 30.As the distance from the camera increases, the point cloud becomes more and more sparse . To counter the effect of reduction in cloud density , we dynamically modify the radius of the search space to get a good density of points for curvature estimation.

We also perform a feature warping with depth to diminish the effect of noise with depth in the features which are calculated from depth, height and curvature. Features are multiplied with a non-linear function in Z, where Z is the depth associated with each pixel. This ensures that colour and unary potentials have more effect on the energy function for point far away from the camera

Figure 5 shows the qualitative improvement in curvature estimation using our approach(on right).Clearly we are able to detect curvature points for longer distancs in comparison to earlier methods(on left).

3.6 Dense Conditional Random Field

We use a fully connected CRF that establishes pairwise potentials on all pairs of pixels in the image. Densely connected graph results in a greatly refined segmentation and labelling, which is crucial for our task. In the fully connected pairwise CRF model, G is the complete graph on X and C is the set of all unary and pairwise cliques. The corresponding Gibbs energy is

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{9}$$

The unary potential $\psi_u(x_i)$ is computed independently for each pixel by a classifier that produces a distribution over the label assignment x_i given image features. The unary potential used in our formulation incorporates shape, texture, location, and color descriptors and is described in Section

Figure 6: We show results for various tough outdoor scenarios. The first column is the left image from stereo camera , second column shows the SegNet results, third column is the final CRF output, forth column shows the curb boundaries by applying canny edge detection on the CRF output

3.2. Since the output of the unary classifier for each pixel is produced independently from the outputs of the classifiers for other pixels, the MAP labelling produced by the unary classifiers alone is generally noisy and inconsistent.

We also perform a temporal update of unary potentials described as follows

$$\psi_u = (P_i^{t-1} + P_i^t)/2 \tag{10}$$

The pairwise potentials in our model have the form

$$\psi_p = \mu(x_i, x_j) \sum_{m=1}^{K} w^{(m)} k^{(m)}(f_i, f_j) \tag{11}$$

The vectors f_i and f_j are feature vectors for pixels i and j in an arbitrary feature space, $w^{(m)}$ are linear combination weights, and μ is a label compatibility function.

$$\begin{aligned} k(f_i, f_j) = w^1 \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right)}_{appearance\ kernel} \\ + w^2 \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2} - \frac{|h_i - h_j|^2}{2\theta_\delta^2}\right)}_{height\ kernel} \\ + w^3 \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|c_i - c_j|^2}{2\theta_\beta^2}\right)}_{curvature\ kernel} \\ + w^4 \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2}\right)}_{smoothness\ kernel} \end{aligned} \tag{12}$$

The appearance and height kernel are inspired by the observation that nearby pixels with similar color and height are likely to be in the same class. The smoothness kernel removes small isolated regions.

A simple label compatibility function is given by the Potts model, $\mu(x_i, x_j) = [x_i \neq x_j]$. It introduces a penalty for nearby similar pixels that are assigned different labels. While this simple model works well in practice, it is insensitive to compatibility between labels. Since we have only two labels, drivable and non-drivable area, this does not affect our formulation.

3.7 Inference

The inference is based on a mean field approximation to the CRF distribution as described in [6]. This approximation yields an iterative message passing algorithm for approximate inference. The key observation presented is that message passing in the presented model can be performed using Gaussian filtering in feature space. This enables us to utilize highly efficient approximations for high-dimensional filtering, which reduce the complexity of message passing from quadratic to linear. This reduction in complexity make this method highly suitable for real-time applications.

3.8 Implementation

We have tested our algorithm on Indian city datasets. The data has been collected using a Point Grey Black Fly cameras, which have been used as a stereo pair. The cameras are setup on our experimental vehicle equipped with a ASUS ROG GR8 computer. Currently all algorithms are run on a Intel quad-core processor. The parameters were estimated by rigorous testing on Indian roads.

4. RESULTS

We show results on dataset collected on our experimental autonomous vehicle with two black fly cameras being used as a stereo pair.

In the paper we show results to verify the superior performance of our method when compared to other existing frameworks. We show more accurate segmentation of the drivable area and hence better curb detection results. Figure 6 and Figure 7 show the qualitative results for various outdoor scenarios, eg straight road, intersection, turns etc.Curb points are detected by the canny edge detection over the output of CRF. We define accuracy as the number of detected curb points that lie inside ground truth(for curb) over the total number of curb points detected by edge detector. This evaluation criteria is best suited for our problem as we are focused mainly in detecting the boundary between

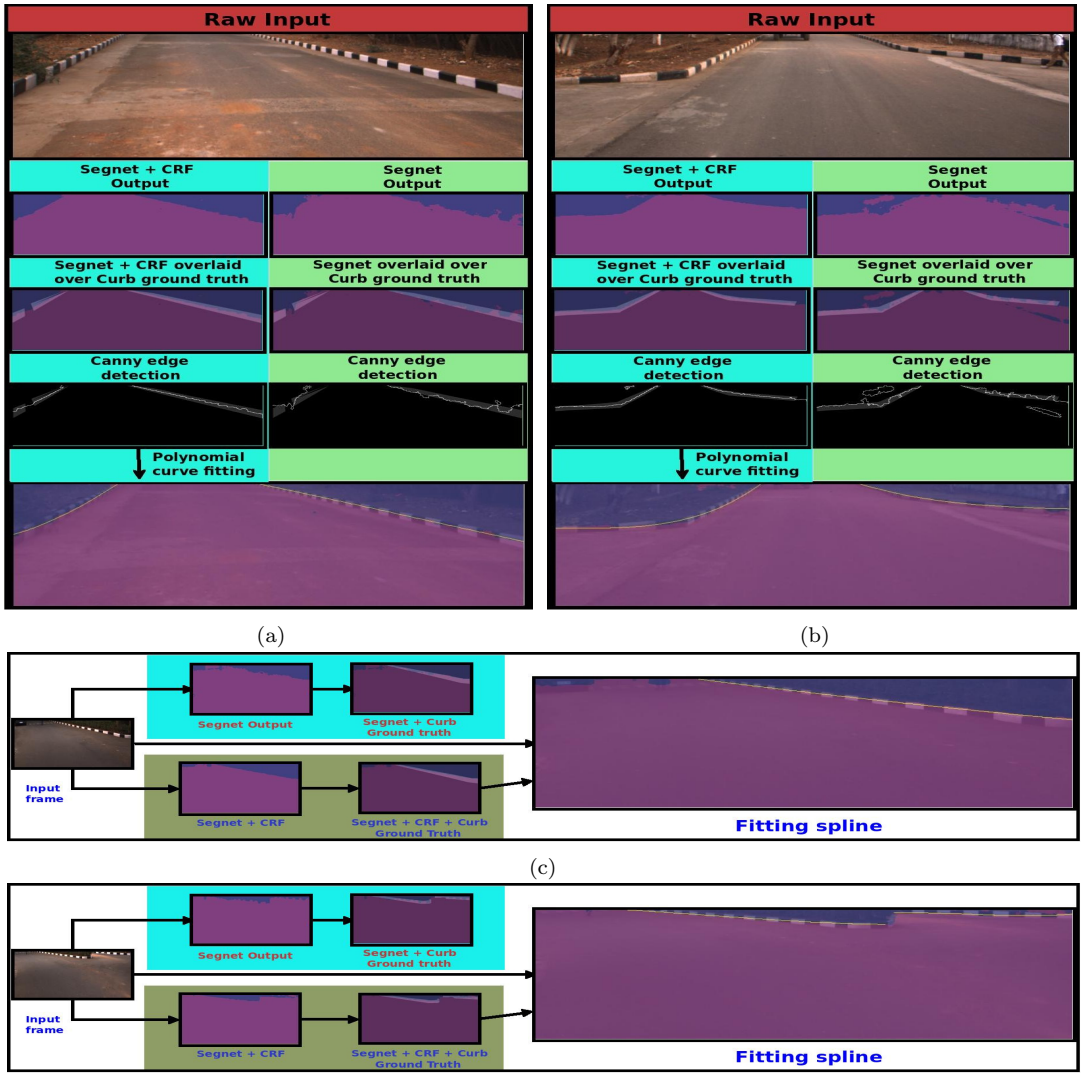

Figure 7: This graphical representation compares our approach's output with the output of SegNet. Figure7(a) shows robustness of our method for curb boundary detection on straight roads. Figure7(b) shows robustness of our method for curb boundary detection on curved roads. Figure7(c) & Figure7(d) shows curb boundary detection in other different scenarios .Spline fitting is done using the output of canny edge detector for better visualization of curb boundaries

Methods	Accuracy
Segnet	71.16874 %
Segnet + CRF(Appearance)	77.80209 %
Segnet + CRF(Appearance + Height + Temporal)	81.03207 %

Table 1: Quantitative results of segnet and our proposed method

the drivable and non drivable area. Following this evaluation criteria, we show the quantitative numbers in comparison to SegNet in the table 1. We also show that temporal smoothness improve the overall classification results. To the best of our knowledge this the first approach where 3D point cloud information has been fused with image information in a dense CRF framework to classify the curb boundaries. We also compare with the feature used in [3] in figure 5. We are able to imrove the range to 40m as compared to 27m in their case.

5. CONCLUSION

Autonomous driving requires accurate segmentation of drivable area. We propose a novel probabilistic framework which fuses the output of neural network with 3D and colour information from a stereo camera pair.With the evaluation criteria mentioned earlier, we show an improvement of nearly 9.8% in curb boundaries detection. The effect of our approach in detection of accurate and smooth curb boundaries is shown in the Figure 7.

In our proposed method, we evaluated effects of incorporating 3D data and temporal filter to improve semantic segmentation results from a pre-trained neural network. It was demonstrated that pre-trained model's results show significant improvement by fusing additional information of the scene. Our probabilistic framework yields better output near edge of the navigable area, improving accuracy of the overall curb boundaries. Usage of dense CRF enables us to estimate probabilities of pixel classes in a more global context of an image. Further we observed that, integrating temporal filter with the result of CRF discards false positives occurring.

6. ACKNOWLEDGMENTS

This work was supported from grants made available by Uurmi Systems.

7. FUTURE WORK

In our future work we also like to include other labels like pedestrian, cars, occlusions etc and jointly optimize for their locations. Also, use the curb detection for better localizing the vehicle, and improve the overall stereo point cloud reconstruction of the road and surroundings. Currently,while segnet and 3d reconstruction modules of our pipeline are real time, both curvature estimation and CRF modules together executes in order of seconds. We will work towards real time implementation of our pipeline.We would like to implement this algorithm on a graphical processing unit(GPU 's) to further accelerate the performance of our algorithms.

8. REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[2] C. Fernández, R. Izquierdo, D. Llorca, and M. Sotelo. Road curb and lanes detection for autonomous driving on urban scenarios. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1964–1969. IEEE, 2014.

[3] C. Fernández, D. Llorca, C. Stiller, and M. Sotelo. Curvature-based curb detection method in urban environments using stereo and laser. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 579–584. IEEE, 2015.

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

[5] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008.

[6] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*, 2011.

[7] F. Oniga and S. Nedevschi. Polynomial curb detection based on dense stereovision for driving assistance. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 1110–1115. IEEE, 2010.

[8] F. Oniga and S. Nedevschi. Processing dense stereo data using elevation maps: Road surface, traffic isle, and obstacle detection. *IEEE Transactions on Vehicular Technology*, 59(3):1172–1182, 2010.

[9] F. Oniga, S. Nedevschi, and M. M. Meinecke. Curb detection based on elevation maps from dense stereo. In *2007 IEEE International Conference on Intelligent Computer Communication and Processing*, pages 119–125. IEEE, 2007.

[10] F. Oniga, S. Nedevschi, and M. M. Meinecke. Curb detection based on a multi-frame persistence map for urban driving scenarios. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 67–72. IEEE, 2008.

[11] C. D. Pantilie and S. Nedevschi. Real-time obstacle detection in complex scenarios using dense stereo vision and optical flow. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 439–444. IEEE, 2010.

[12] M. Pauly, M. Gross, and L. P. Kobbelt. Efficient simplification of point-sampled surfaces. In *Visualization, 2002. VIS 2002. IEEE*, pages 163–170, 2002.

[13] S. Se and M. Brady. Vision-based detection of kerbs and steps. In *BMVC*, 1997.

[14] J. Siegemund, U. Franke, and W. Förstner. A temporal filter approach for detection and reconstruction of curbs and road surfaces based on conditional random fields. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 637–642. IEEE, 2011.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.