FEATURE++: CROSS DIMENSION FEATURE FUSION FOR ROAD DETECTION

*Wenli He[1], Guorong Cai[2], Zhun Zhong[1], and Songzhi Su[1]**

[1] Cognitive Science Department, Xiamen University, Xiamen, Fujian, 361005, China
[2] Computer Engineering College, Jimei University, Xiamen, Fujian, 361021, China

ABSTRACT

Road detection is a key component of Advanced Driving Assistance Systems, which provides valid space and candidate regions of objects for vehicles. Mainstream road detection methods have focused on extracting discriminative features. In this paper, we propose a robust feature fusion framework, called "Feature++", which is combined with superpixel feature and 3D feature extracted from stereo images. Then a neural network classifier is been trained to decide whether a superpixel is road region or not. Finally, the classified results are further refined by conditional random field. Experiments conducted on the KITTI ROAD benchmark show that the proposed "Feature++" method outperforms most manually designed features, and are comparable with state-of-the-art methods that based on deep learning architecture.

Index Terms— Road detection, Kernel descriptors, Gabor, Multi-feature fusion, Conditional Random Field

1. INTRODUCTION

Advanced Driving Assistance Systems (ADAS) have been receiving considerable attention over the past few years, due to the potential improvement of traffic efficiency and human safety. In the pipeline of ADAS, road detection plays an important role. Current road detection methods are based on different sensors, such as monocular camera, stereo camera, laser range finders or fusion of them.

Among the sensors mentioned before, monocular camera is the cheapest while the visible image provides rich color information. As for monocular road feature descriptor, low level cues have been widely used, such as illumination invariant intrinsic images [1][2], combination of color planes and texture [3]; In order to improve the performance under the situation of unmarked road, spatial ray feature [6] and contextual feature[7] are proposed; Noteworthy that state-of-the-art methods use Convolutional Neural Networks for feature learning such as [8]. Compared to manually designed features, learned features are more discriminative while requiring more computational cost. As for classifier, traditional machine learning methods, such as, SVM [9], Gaussian mixture model [10], boosting [11] and Artificial Neural Network [12] have been used in road detection task. When the 2D feature is not discriminative between road and non-road areas such as the similarly paved side-walks and road, detection may fail. Besides, due to the lack of 3D information, monocular based methods are sensible to illumination conditions such as shadows, and back-lighting.

LIDARs based road detection methods generally analyze the 3D scene and take the plat areas as road candidate. For instance, Thrun [13] proposes a min-max elevation map, Moosmann [14] uses a local convexity criterion, and Chen [15] uses Gaussian Process Regression in the polar grid map to detect road. However, due to lacking of color and texture information, it cannot distinguish the road and non-road areas which have little difference in height. Moreover, the point clouds are always too sparse for segmenting road area.

Among all sensors, stereo camera, which is at an affordable price while covers the virtues of monocular cameras and LIDARs, is mainly concerned in recent researches [16][17]. As for stereo based road detection approaches, the disparity or the 3D information are usually employed, such as the V-disparity [18] map, stochastic occupancy grid [19], digital elevation map [20], and cubic B-spline curve fitting [21]. Since the stereo matching would get error more or less, direct use of the disparity and 3D information may lead to unreliable results. To reduce the influence of stereo error, Berneshawi [22] combines simple 2D and 3D information to segment the road area. However, the simple color statistic cannot represent texture information well. In addition, the simple 2D and 3D statistic would cause accuracy loss.

In this paper, motivated by NNP framework [22], we try to process multi-feature fusion to acquire more appropriate road feature. We partition the image into superpixels, and road features are extracted from these segments to learn a three-layer artificial neural network (ANN) as road classifier. Then conditional random field (CRF) is used to refine the classified results. In the proposed framework, we find that the Gabor feature [23] is particularly appropriate for texture representation and discrimination. Moreover, the kernel descriptors [24] provide a unified framework to turn pixel attributes into patch-level features with low loss of accuracy. Therefore, we fuse the Gabor feature, kernel descriptors, some other simple color and 3D spatial information for road detection task.

*Corresponding author.

2. PROPOSED METHOD

The pipeline of our method is depicted as Fig. 1. Firstly, we use the method [25] which jointly processes image segmentation and stereo matching to partition the segment image into superpixels and get disparity map. Secondly, 2D and 3D road features, such as Gabor feature, Kernel descriptors, and 3D spatial information, have been combined as the descriptor of superpixels. Then we perform multi-feature fusion via ANN classifier. Due to the fact that ANN classifier does not take the contextual information into consideration, we use CRF [26] to refine the output of the classifier.

2.1. Multi-feature fusion

In order to better reflect the synergies of our 2D and 3D features, we following the idea of [28], which recommends that features can be fused in the early stage. Feature vectors extracted from both 2D and 3D information are combined as the input of the ANN classifier:

$$v_{fusion} = (v_{2D}, v_{3D}) \quad (1)$$

Where v_{2D} and v_{3D} are 2D and 3D features extracted from superpixel and depth maps, respectively. Typical 2D features are Gabor feature, gradient kernel descriptor, and 3D features include spin kernel descriptors, plane feature, depth gradient kernel descriptors, etc. In order to choose an appropriate cross feature fusion strategy, we first evaluate the efficiency of 2D and 3D features independently. Then selectively choose the 2D combination and 3D combination for accuracy evaluation. The detail of feature selection and combination will be shown in the section of experiments.

As for Gabor transform [23], which extracts feature with different frequencies and orientation, it is particularly appropriate for texture representation. Therefore, in the road detection task, we define 32 Gabor filters for image convolution: 4 frequencies and 8 orientations each frequency. Each pixel gets 32 Gabor results and then we average them over superpixel to get a 32-dimensional vector as road feature.

As for Kernel descriptors (KDES) [24], it proposes a unified framework, which turns pixel attributes into patch-level features in the kernel view with low loss of accuracy. For road detection task, since gradient, color, depth gradient, and spin cover the image variations, appearance, depth variations, and surface normal information respectively, we then evaluate four KDES, including gradient (GKDES) [24], color (RGBKDES) [24], depth gradient (DGKDES) [24], and spin (SPINKDES) [27]. We average each KDES over superpixel as road feature.

2.2. ANN for cross feature classification

Note that the fused feature includes 2D and 3D information. It's fairly challenging to trade off the importance of each dimension in a vector. Since artificial neural network has the advantages of adaptability and generalization, it is an efficient tool for feature fusion and classification. In the proposed method, we use a three-layer network, which consists of the input, hidden and output layers. In our experiments, we set the number of hidden layer neurons equal to the dimension of input feature. Moreover, in order to show the feasibility of the proposed ANN strategy, we also compare the performance of ANN with SVM in the section 3. The prediction is given as the following equation:

$$p(v_{fusion}) = \sigma(\omega_o \sigma(\omega_h \cdot v_{fusion})) \quad (2)$$

where v_{fusion} is the input feature, ω_o and ω_h are the weight matrices of the output and hidden layer respectively, and σ is the sigmoid function. The predicted probabilities represent the road confidences and CRF will take these confidences as input.

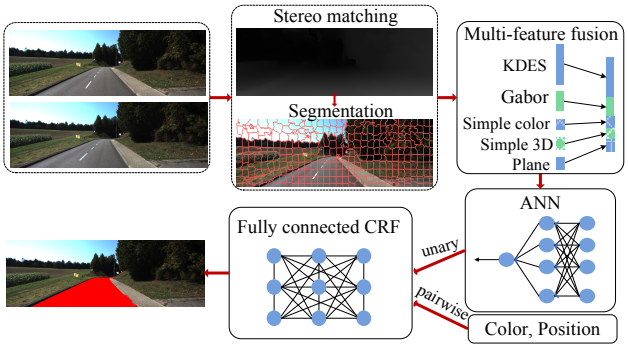


Fig.1. The pipeline of the proposed road detection algorithm

2.3. CRF for road detection

CRF is an efficient tool for multi-class image segmentation and labeling [29]. Note that road detection is a two-class (road and non-road) labeling problem. We can use a fully connected CRF model [26] to refine the classified results. Consider a random field I defined over variables $\{I_1, \cdots, I_N\}$, where N is the size of input image and I_j is the color vector of pixel j. Similarly, consider X defined over variables $\{X_1, \cdots, X_N\}$, where X_j is the road label assigned to pixel j. Then the posterior probability of the overall labeling given the observed image I can be expressed as:

$$P(X|I) = \frac{1}{Z(I)} exp\left(-\sum_{c \in C_\varsigma} \phi_c(X_c|I)\right) \quad (3)$$

where C_ς is the set of all cliques, ϕ_c is the potential function of clique c, and Z is used to normalize. Then the most probable road label of image I can be written as:

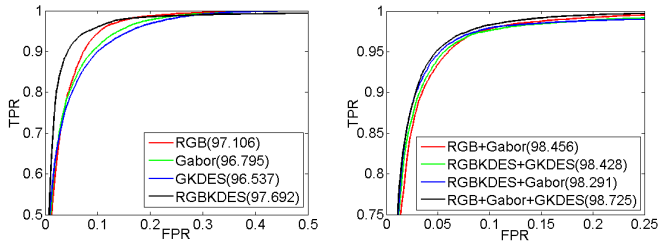
$$x^* = argmin_x - log\big(P(X|I)\big) = argmin_x E(x) \quad (4)$$

in which $E(x)$ is the corresponding Gibbs energy:

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \quad (5)$$

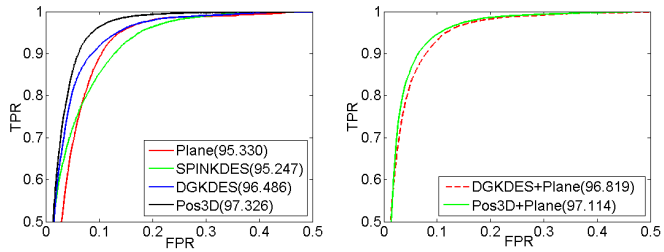
The unary potential $\psi_u(x_i)$ takes the negative log-likelihood of the road confidence in the pixel i which is predicted by learned ANN. The pairwise potentials use the form as:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) k(f_i, f_j) \quad (6)$$



(a) 2D features (b) Combination of 2D features

Fig.2. Evaluation of 2D features.



(a) 3D features (b) Combination of 3D features

Fig.3. Evaluation of 3D features

where $\mu(x_i, x_j) = [x_i \neq x_j]$ and $k(f_i, f_j)$ is defined by the color vectors I_i, I_j and position p_i, p_j:

$$k(f_i, f_j) = \omega^{(1)} exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) + \omega^{(2)} exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) \quad (7)$$

in which ω are linear combination weights and θ are parameters controlling the degrees of pixel nearness and similarity. As for our road detection method, we set $\omega^{(1)}$ and $\omega^{(2)}$ as 10, θ_γ as [5, 5], θ_α as [90, 90], and θ_β as [5, 5, 5]. In the equation (7), the first part helps nearby pixels with similar color be likely in the same class, and the latter part helps remove small isolated regions. We evaluate the necessity of CRF in the experiments section.

3. EXPERIMENTS

3.1. Datasets

We use the KITTI ROAD benchmark [30] to evaluate our approach. It contains 289 annotated image pairs for training and 290 pairs for testing. Both sets consist of three categories of road scenes: urban unmarked (UU), urban marked (UM), and urban multiple marked lanes (UMM). We do not make distinction between the three categories. Methods are ranked according to the pixel-wise maximum F-measure on the Bird's-eye view (BEV) space. To select the most adequate parameters and evaluate the different feature fusion strategies, we divide the training set into two sets: 216 image pairs for training and 73 for validation.

3.2. Features Evaluation

We first evaluate the 2D and 3D features respectively, then selectively fuse them to get the best feature fusion.

The 2D features we use include RGB, Gabor, GKDES, and RGBKDES, where RGB represents the mean and standard deviation of each channel. We compare the representation ability of them, and as shown in Fig. 2(a), RGBKDES is the most discriminative for road. In addition, due to that RGBKDES and RGB represent color information, Gabor feature provides texture and GKDES describes gradient, according to the characteristics of the road, we process them with different combination. As shown in Fig. 2(b), we compare four combinations: "RGB + Gabor", "RGBKDES + GKDES", "RGBKDES + Gabor" and "RGB + Gabor + GKDES". The results show that the "RGB + Gabor + GKDES" performs the best.

The 3D features we choose to evaluate consist of Plane, SPINKDES, DGKDES, and Pos3D. The Plane represents the angles and inliers percentage of plane which is fitted in a superpixel, and Pos3D is the mean and deviation of 3 coordinates. As can be seen in the Fig. 3 (a), Pos3D is the most discriminative for road. As we all know that both SPINKDES and Plane represent the angle information, and SPINKDES computes more details. Due to that the 3D information which gotten by stereo matching would have some error more or less, the SPINKDES, more affected by stereo noise, performs not as well as Plane. Moreover, according to that both Pos3D and DGKDES represent the information of 3D position changes, we process two fusions for road: "DGKDES + Plane" and "Pos3D + Plane". As shown in Fig. 3(b), "Pos3D + Plane" performs better.

Finally, we fuse the 2D and 3D features to put them work together. According to the above experiments, as shown in Fig. 4, we evaluate three combinations: "RGBKDES + Gabor + DGKDES + Plane", "RGB + Gabor + GKDES + Pos3D + Plane" and "RGB + Gabor + Pos3D + Plane". We select the best feature fusion "RGB + Gabor + GKDES + Pos3D + Plane" as our road feature. In addition, as shown in Fig. 4, in accordance with the comparison of monocular ("RGB + Gabor + GKDES") and stereo performance, we can see the importance for the use of stereo information.

3.3. Classifier evaluation

We also evaluate how the classifier affects the performance. For this purpose, we assess a SVM classifier as comparison using the LIBSVM library. The SVM type is set as "C-SVC" and the kernel function is RBF (radial basis function). Fig. 5 shows that ANN outperforms SVM in our method.

3.4. Benchmark Submission

We submit the results of our method, which uses the "RGB + Gabor + GKDES + Pos3D + Plane" as road feature and ANN as the classifier, to the KITTI-ROAD Benchmark. To verify the necessity of the CRF post-process, as shown in

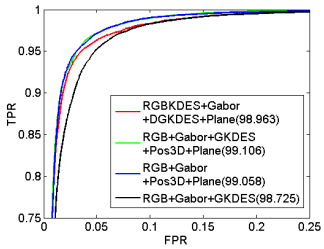


Fig. 4. Evaluation of 2D and 3D feature fusions

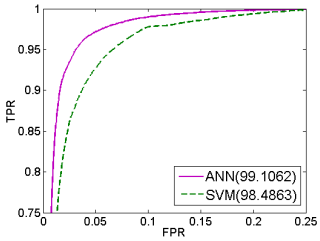


Fig. 5. Evaluation of SVM and ANN

the Table 1, we submit two results: without CRF and with CRF. The CRF can help the performance improve 2-3%.

Moreover, we compare the results of three road scenes. We know that the most unmarked roads are in the rural areas and the surrounded environments are relatively complex. Trees and buildings on both sides of the narrow road usually cast shadows which would disturb the detection. Therefore, as can be seen in Table 1, the performance in unmarked roads is not as good as in marked roads.

Table 2 shows the comparison of our method with others. BL [30] provides a low bound for the performance that any road detection algorithm should achieve; CB [7] proposes a contextual feature; SRF [31] presents a structured random forest-based road detection algorithm. BL, CB and SRF are monocular based; FusedCRF [11] fuses the LIDAR and monocular image to detect road; NNP, which is based on stereo, is basic framework of our method. We can see that our method performs the best. Fig. 6 shows a visual comparison of our method, NNP, and FusedCRF.

4. CONCLUSIONS

In this paper, we have proposed a robust feature for road detection, which selectively fuses the Gabor, kernel descriptors, simple color and 3D spatial information. Meanwhile, we have evaluated each component of our system. Experiments show that our method outperforms the most of methods which use the manually designed features on the KITTI ROAD benchmark.

However, Illumination conditions usually affect the performance, especially in unmarked road scenes. In our future works, illumination invariant features would be included in the road feature fusion and used in the pairwise potentials of CRF.

Table 1: Results [%] of our method with or without CRF on the KITTI Road testing dataset.

With CRF						
Benchmark	**MaxF**	**AP**	**PRE**	**REC**	**FPR**	**FNR**
UMM_ROAD	93.55	92.34	92.77	94.34	8.08	5.66
UM_ROAD	90.43	88.83	89.63	91.26	4.81	8.74
UU_ROAD	87.40	85.48	86.19	88.64	4.63	11.36
URBAN_ROAD	**91.12**	**89.51**	**90.16**	**92.10**	**5.54**	**7.90**
Without CRF						
URBAN_ROAD	**89.48**	**90.45**	**87.74**	**91.28**	**7.02**	**8.72**

Table 2: Methods results [%] comparison on the KITTI Road Benchmark

URBAN - BEV space						
Method	**MaxF**	**AP**	**PRE**	**REC**	**FPR**	**FNR**
Feature++	**91.12**	**89.51**	**90.16**	**92.10**	**5.54**	7.90
NNP[22]	89.68	86.50	89.67	89.68	5.69	10.32
CB[7]	88.97	79.69	89.50	88.44	5.71	11.56
FusedCRF[11]	88.25	79.24	83.62	93.44	10.08	**6.56**
SRF[31]	82.44	87.37	80.60	84.36	11.18	15.64
BL[30]	75.89	79.28	71.56	80.77	5.65	19.23

5. ACKNOWLEDGMENTS

This work is supported by the Nature Science Foundation of China (No. 61572409, No.61402386 & No. 61571188), Fujian Province 2011 Collaborative Innovation Center of TCM Health Management and Collaborative Innovation Center of Chinese Oolong Tea Industry-Collaborative Innovation Center (2011) of Fujian Province.

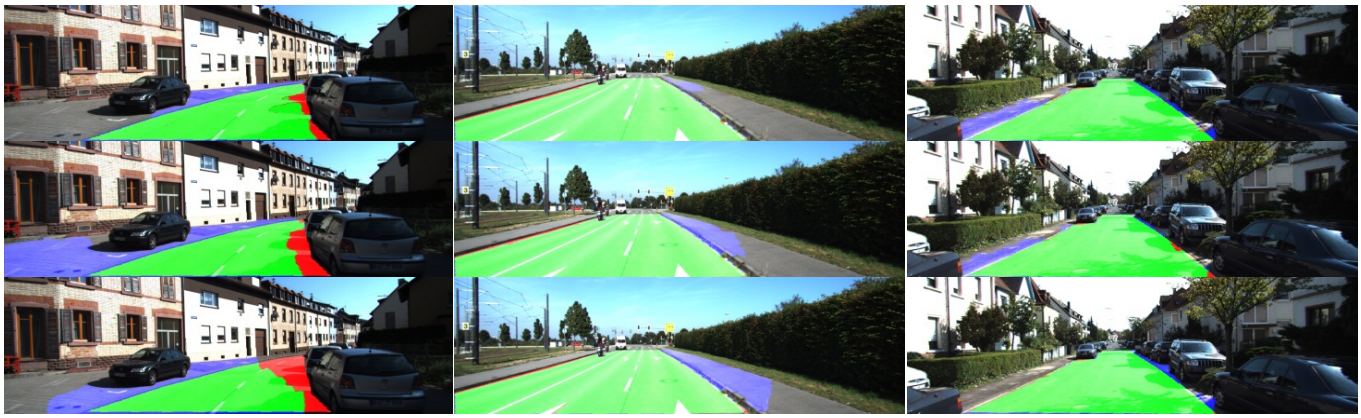


Fig.6. Sample results in the perspective image. Red denotes false negatives, blue areas correspond to false positives and green represents true positives (Top to bottom : Ours, NNP, FusedCRF. Left to right: UM, UMM, UU)

6. REFERENCES

[1] J. Alvarez and A. M. Lopez, "Road detection based on illuminant invariance," I*ntelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 1, pp. 184–193, 2011.

[2] B. Wang, V. Fremont, and S. Rodriguez, "Color-based road detection and its evaluation on the KITTI road benchmark," *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pp. 31–36, June 2014.

[3] J.M. Alvarez, T. Gevers, Y. Lecun and A.M. Lopez, "Road scene segmentation from a single image," *ECCV 2012 Ser. Lecture Notes in Computer Science*, vol. 7578, pp. 376-389, 2012.

[4] C. Rasmussen, "Grouping dominant orientations for ill-structured road following," in *Proc. IEEE Comp. Soc. Conf. Computer Vision and Pattern Recognition*, vol. 1, July 2004, pp. 470–477.

[5] H. Kong, J.-Y. Audibert and J. Ponce, "Vanishing point detection for road detection," in *CVPR 2009, IEEE Computer Society*, pp. 96-103, 2009.

[6] J. Fritsch, T. Kuehnl and F. Kummert, "Monocular road terrain detection by combining visual and spatial information," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1-11, 2014.

[7] C. Mendes, V. Frémont and D. Wolf, "Vision-Based Road Detection using Contextual Blocks," *ArXive-prints, 2015. [Online].Available:http://arxiv.org/abs/1509.01122*

[8] R. Mohan, "Deep Deconvolutional Networks for Scene Parsing," *arXiv preprint arXiv:1411.4101, 2014.*

[9] Y. Alon, A. Ferencz and A. Shashua, "Off-road path following using region classification and geometric projection constraints", *Proc. IEEE Conf. Computer Vision Pattern Recognition 2006*, pp. 689-696.

[10] H. Dahlkamp, A. Kaehler, D. Stavens, S.Thrun, and G. R. Bradski, "Self-supervised monocular road detection in desert terrain," in *Robot. Sci. Syst. Conf. (RSS)*, 2006.

[11] Liang Xiao, Bin Dai, Daxue Liu, Tingbo Hu, and Tao Wu. "CRF based road detection with multi-sensor Fusion," *In Intelligent Vehicles Symposium (IV), 2015 IEEE*, pp. 192–198, 2015.

[12] G. Vitor, D. Lima, A. Victorino and J. Ferreira, "A 2D/3D Vision Based Approach Applied to Road Detection in Urban Environments," *Intelligent Vehicles Symposium (IV), 2013 IEEE*, pp. 952-957, 2013.

[13] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron,J. Diebel, P. Fong, J. Gale, M. Halpenny, and G. Hoffmann, et al., "Stanley: The robot that won the darpa grand challenge," in *The 2005 DARPA Grand Challenge. Springer*, 2007, pp. 1–43.

[14] F. Moosmann, O. Pink and C. Stiller, "Segmentation of 3D lidar data in non-flat urban environments using a local convexity criterion," *IEEE Intelligent Vehicles Symposium*, pp. 215-220, 2009.

[15] T. Chen, B. Dai, R. Wang, and D. Liu, "Gaussian-process-based realtime ground segmentation for autonomous land vehicles," *Journal of Intelligent & Robotic Systems (JINT)*, vol. 76, pp. 563–582, Sep 2013.

[16] Zhong Z, Su S, Cao D, and Li S, "Detecting Ground Control Points via Convolutional Neural Network for Stereo Matching," arXiv preprint arXiv:1605.02289, 2016.

[17] Zhong Z, Lei M, Li S and Fan J, "Re-ranking Object Proposals for Object Detection in Automatic Driving," arXiv preprint arXiv:1605.05904, 2016.

[18] R. Labayrade, D. Aubert and J. Tarel, "Real Time Obstacle Detection in Stereovision on Non Flat Road Geometry through 'v-Disparity' Representation," *Proc. IEEE Intelligent Vehicles Symp*, vol. 2, pp. 17-21.

[19] H. Badino, W. Franke and R. Mester, "Free Space Computation Using Stochastic Occupancy Grids and Dynamic Programming," *Proc. Int'l Conf. Computer Vision Workshop Dynamical Vision*.

[20] F. Oniga and S. Nedevschi, "Processing dense stereo data using elevation maps: Road surface, traffic isle, and obstacle detection," *IEEE Trans. Veh. Technol*, vol. 59, no. 3, pp.1172 -1182, March 2010.

[21] J. K. Suhr, and H. G. Jung. "Dense Stereo-Based Robust Vertical Road Profile Estimation Using Hough Transform and Dynamic Programming," *IEEE Transactions on Intelligent Transportation Systems*, online published on 4 Dec. 2014.

[22] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler and R. Urtasun,"3D Object Proposals for Accurate Object Class Detection," *NIPS*, 2015.

[23] A.G. Ramakrishnan, S. Kumar Raja and H.V. Raghu Ram, "Neural network-based segmentation of textures using Gabor features," *Proc. 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 365 - 374, 2002.

[24] L. Bo, X. Ren and D. Fox, "Kernel descriptors for visual recognition," *NIPS*, 2010.

[25] Yamaguchi, K.,McAllester, D., and Urtasun, R., " Efficient joint segmentation, occlusion labeling, stereo and flow estimation," In *Computer Vision–ECCV 2014, Springer*, pages 756–771.

[26] K, Philipp and K, Vladlen, "Efficient inference in fully connected CRFs with Gaussian edge potentials," In *NIPS*, 2011.

[27] L. Bo , X. Ren and D. Fox , "Depth kernel descriptors for object recognition," *Proc. Int. Conf. IROS*, pp. 821-826, 2011

[28] D. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[29] J. Shotton, J. Winn, C. Rother and A. Criminisi, "Textonboost: Joint appearance shape and context modeling for multi-class object recognition and segmentation," *ECCV* 2006, pp. I:1-15, 2006.

[30] J.Fritsch, T.Kuehnl and A.Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," *Proc. IEEE Intelligent Transportation Systems*, pp. 1693-1700, 2013.

[31] Xiao L, Dai B, Liu D, Zhao D, Wu T. "Monocular Road Detection Using Structured Random Forest," *Int JAdv Robot Syst*, 2016, 13(2016):101.