

LETTER

Ground Plane Detection with a New Local Disparity Texture Descriptor

Kangru WANG^{†,††a)}, Lei QU[†], Lili CHEN[†], *Nonmembers*, Jiamao LI[†], *Member*,
Yuzhang GU[†], Dongchen ZHU[†], and Xiaolin ZHANG[†], *Nonmembers*

SUMMARY In this paper, a novel approach is proposed for stereo vision-based ground plane detection at superpixel-level, which is implemented by employing a Disparity Texture Map in a convolution neural network architecture. In particular, the Disparity Texture Map is calculated with a new Local Disparity Texture Descriptor (LDTD). The experimental results demonstrate our superior performance in KITTI dataset.

key words: ground plane detection, Local Disparity Texture Descriptor (LDTD), stereo vision, convolution neural network

1. Introduction

Ground plane detection is a key component of Driver Assistance Systems, which helps to improve the traffic safety and efficiency with drivable space information. Furthermore, ground plane detection is widely exploited in many applications, such as object detection [1], [2] and free space estimation [3], [4].

The approach of stereo vision-based ground plane detection has two main trends: Euclidian space-based and disparity space-based. The algorithm proposed by [5] uses the RANSAC-based approach to fit a ground plane with the 3D points. However, it can not cope with non-flat ground plane. B-spline curve [4], [6], [7], polynomial function [8], [9] and piece-wise linear function [10] were proposed to model the ground plane, which can address the varying longitudinal slope but ignore the latitudinal slope. Compared to those Euclidian space-based approaches, the approach based on disparity space has less computational cost without calculating the 3D point location. Labayrade et al. [11] first proposed the well-known V-disparity algorithm. The algorithm accumulates the pixels with same disparity value along image rows to form the V-disparity map where the ground plane is projected as slanted lines. Thus, ground plane detection is simplified to line extraction. However, it has difficulties in detecting the ground plane with latitudinal slopes. The sub-V-disparity algorithm proposed by [12] copes with the latitudinal slope by a sliding windows paradigm. This algorithm divides the disparity map into several windows

where the ground is considered flat in latitudinal direction. However, it is hard to define the size of the sliding window. If the window size is too small there are insufficient ground pixels to be projected in the sub-V-disparity map, whereas the ground is not considered flat in latitudinal direction within large window. In addition, the above V-disparity and sub-V-disparity algorithms are performed by estimating a global road profile within a whole image or a window, which have the limitation in representing the detailed properties of ground plane and keeping resolution in distant area.

This paper proposes a new approach to address above mentioned limitations in ground plane detection. Inspired by the semantic segmentation approach [13], the proposed approach is performed at superpixel-level. First, the Disparity Texture Map is calculated with the new proposed LDTD from the disparity map and then divided into superpixels. Then, the ground plane is detected by classifying the superpixels into ground or non-ground using the proposed convolution neural network, while contextual information is employed to improve the classification accuracy. The main contributions are presented as follows: (1) A novel approach for ground plane detection is proposed by integrating the new LDTD with a convolution neural network and superpixel segmentation, which can address most of terrains and perform precise segmentation of ground plane. (2) The new LDTD is proposed to extract the feature of ground plane only requiring disparity information, which is suitable for flat/non-flat ground plane and multi-ground plane. (3) The contextual information is utilized to improve the classification accuracy of the convolution neural network.

2. Proposed Approach

The overview of our approach is illustrated in Fig. 1. The Disparity Texture Map (Fig. 1 (b)) is computed from the disparity map (Fig. 1 (a)) with the LDTD. Then, the Disparity Texture Map is segmented into superpixels (Fig. 1 (e)) based on the color information of the left stereo image (Fig. 1 (c)). To detect the ground plane region, the patch centered around each superpixel is extracted and inputted into the convolution neural network. The output of the network is the class of each considered superpixel. Details are described in the following subsections.

Manuscript received March 7, 2017.

Manuscript revised June 4, 2017.

Manuscript publicized June 27, 2017.

[†]The authors are with Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China.

^{††}The author is with University of Chinese Academy of Sciences, Beijing, 100049, China.

a) E-mail: wangkangru@mail.sim.ac.cn

DOI: 10.1587/transinf.2017EDL8053

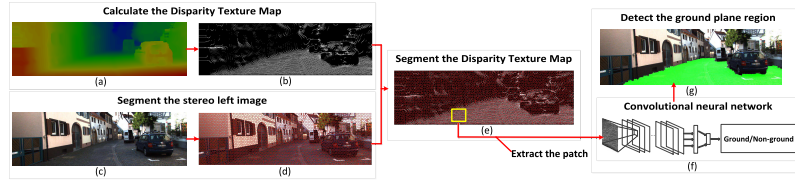


Fig. 1 Our algorithm flowchart. (a) is the disparity map and (b) is the Disparity Texture Map estimated from (a) with the proposed LDTD. (c) is the stereo left image and (d) is the segmented result of (c) utilizing the SLIC algorithm. (e) is the segmented Disparity Texture Map based on (d). The patch centered around each superpixel of (e) is extracted and inputted into the convolution neural network (f). The output of the network (f) is the class of the considered superpixel. (g) is the result of ground plane detection. The color in the disparity map (a) encodes disparity value, red means large and blue represents small.

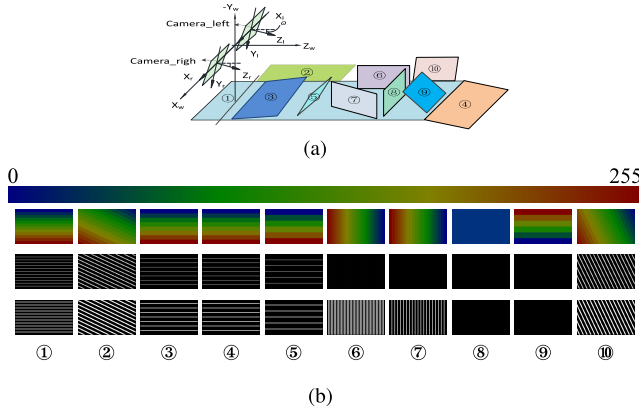


Fig. 2 Examples of the Disparity Texture Map. (a) Typical planes in world coordinate system. (b) The ten images in the first row are the disparity maps of the planes in figure (a). The ten images in the second row are the corresponding Disparity Texture Maps. The ten images in the third row are the binary maps of the Disparity Texture Maps for visualization. The labels are consistent with the planes in figure (a). The color in the disparity maps encodes the disparity value. There are notable differences between the ground plane (plane ①, ②, ③, ④) and obstacle plane (⑤, ⑥, ⑦, ⑧, ⑨, ⑩) in the Disparity Texture Maps.

2.1 Calculate the Disparity Texture Map

Ten typical planes in the world coordinate system are illustrated in Fig. 2(a) and their disparity maps are shown in Fig. 2(b). The plane ① represents the horizontal ground plane. The plane ② represents the ground plane with latitudinal slope while the plane ③ and ④ represent the ground plane with longitudinal slope. The plane ⑤, ⑥, ⑦, ⑧, ⑨ and ⑩ represent the typical planes of obstacle.

Based on the disparity characteristics of the ground plane and obstacle, we define the LDTD as a structure, which consists of 3x3 sub-blocks, as shown in Fig. 3. To calculate the Disparity Texture Map, we employ the LDTD to scan the disparity map along each coordinate with a stride of 1 pixel. The value of each pixel in the Disparity Texture Map is formulated as:

$$V = \sum_{i=1}^8 C_i 2^{8-i} \quad (1)$$

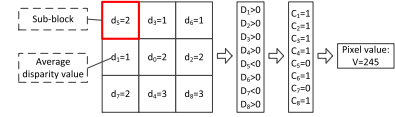


Fig. 3 An example of the LDTD operation process. The LDTD is defined as a structure which consists of 3x3 sub-blocks. d_i is the average disparity value of the corresponding sub-block. The output value V is assigned to the pixel in the center of the LDTD structure.

where C_i is defined as:

$$C_i = \begin{cases} 1, & \text{if } D_i > 0; \\ 0, & \text{else.} \end{cases} \quad (2)$$

Where D_i is defined as:

$$D_1 = (d_0 + d_1 + d_2)/3 - (d_3 + d_5 + d_6)/3; \quad (3)$$

$$D_2 = (d_4 + d_7 + d_8)/3 - (d_0 + d_1 + d_2)/3; \quad (4)$$

$$D_i = d_0 - d_i \quad (i = 3, 5, 6); \quad (5)$$

$$D_i = d_i - d_0 \quad (i = 4, 7, 8); \quad (6)$$

Where d_i is the average disparity value of the corresponding sub-block region, as shown in Fig. 3.

Disparity Texture Maps of the typical planes described above are illustrated in Fig. 2(b). For the ground plane, the texture is dense with high gray value, while the texture direction is transverse (plane ①, ③, ④) or with small slope angle (plane ②). The texture corresponding to the obstacle plane ⑤ is sparser than ground plane ①, ③, ④. The texture corresponding to the obstacle plane ⑩ has larger slope angle than ground plane ②. The plane ⑥, ⑦, ⑧, ⑨ have notable difference from ground plane in texture direction and gray value. In addition, the LDTD is robust to most of terrain variation since the common non-flat ground plane or multi-ground plane can be considered as a combination of the typical ground plane described above.

2.2 Segment the Disparity Texture Map

To further improve the segmentation precision of ground plane, the Disparity Texture Map is segmented into superpixels as shown in Fig. 1(e). The superpixel is extracted using the SLIC (Simple linear iterative clustering) algorithm [14] based on the color information of the stereo left

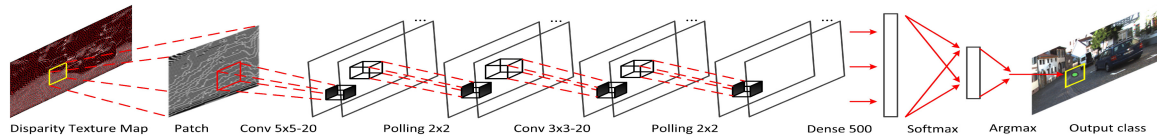


Fig. 4 Our network architecture. The input of the network is the patch extracted from the Disparity Texture Map and the output is the class of the superpixel in the center of the patch. The first fully connected layer has 500 neurons and the final layer has 2 neurons implementing the Softmax loss function.

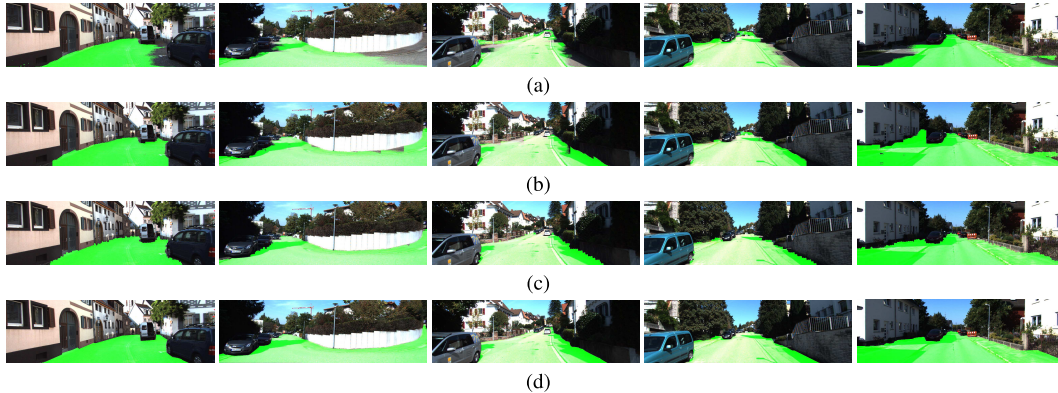


Fig. 5 Qualitative results of ground plane detection. (a) V-disparity. (b) Sub-V-disparity. (c) Our method. (d) Ground truth.

image. The SLIC algorithm has the advantage of excellent boundary adherence, which is of great importance to the ground plane segmentation. The pixels of one superpixel can be considered belonging to the same object since they are clustered based on similar color and spatial information. Thus, each superpixel is analyzed as a unit in the following classification step. In our paper, about 4000 superpixels are generated in each image.

2.3 Detect the Ground Plane Region

The ground plane is detected by classifying each superpixel in the Disparity Texture Map using the proposed convolution neural network, which is inspired by the LeNet-5 [15]. The network architecture is illustrated in Fig. 4. To improve the classification accuracy, the contextual information is employed. Inspired by [16], the input of the network is a relatively large patch while the output is the class of the superpixel in the center of the patch. Each convolutional layer and the first fully connected layer are all followed by ReLU activation. A dropout rate of 50% is employed in the first fully connected layer.

3. Experimental Result

3.1 Experimental Data

Our approach is evaluated on the datasets from the KITTI benchmark [17]. The training dataset consists of 122 images from 7 sequences (09_26_d_09, 09_26_d_18, 09_26_d_35, 09_26_d_48, 09_28_d_34, 09_28_d_45, 09_28_d_68). The testing dataset consists of 303 images from other 10 se-

quences (09_26_d_17, 09_26_d_39, 09_26_d_61, 09_26_d_64, 09_26_d_86, 09_26_d_93, 09_28_d_66, 09_30_d_33, 10_03_d_27, 10_03_d_47). These sequences contain different scenes with flat/non-flat ground plane and multi-ground plane. Ground truth is labeled manually and disparity maps are estimated by the algorithm of Zhontar et al. [18].

3.2 Training Scheme

The samples are created by extracting the patches centered around superpixels in the Disparity Texture Map of the training dataset. The sample is labeled as ground/non-ground if its corresponding superpixel contains more than 90% pixels belonging to ground/non-ground region. 25% of these samples are chosen for validation and the others for training. Our work is implemented using the Caffe framework [19]. The network is optimized using stochastic gradient descent, while the initial learning rate is 0.01, momentum is 0.9, weight decay rate is 0.0001 and gamma is 0.1. The batch size is 64. The training ends after 30 fully epochs.

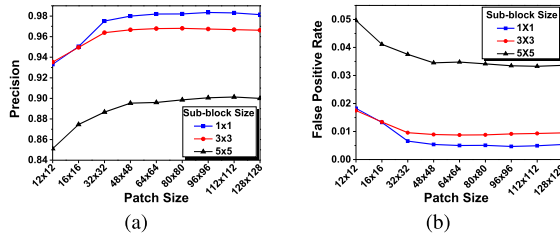
3.3 Evaluation Result

We evaluate our approach with two classical methods: V-disparity [11] and sub-V-disparity [12], which are also implemented in disparity space.

Figure 5 shows the qualitative results. As we can see, the wrongly predicted ground near the car wheels in the first image of Fig. 5 (b) is detected correctly in the first image of Fig. 5 (c). The false predictions in distant area in the fourth image of Figs. 5 (a), (b) are correctly detected in the fourth image of Fig. 5 (c). Our approach provides more precise seg-

Table 1 Comparison on ground plane detection.

	F_1 -score	Precision	Recall	FPR	Accuracy
V-disparity [11]	87.70%	92.34%	84.77%	2.26%	94.55%
Sub-V-disparity [12]	92.98%	92.47%	93.78%	2.32%	96.80%
Our method	96.86%	98.36%	95.45%	0.47%	98.57%

**Fig. 6** The performance of our approach with different patch size and sub-block size. (a) The precision of our approach. (b) The FPR of our approach. Note that the 12x12 patch size is approximate to the case without any contextual information.

mentation of the ground plane than the baselines.

The quantitative evaluation results are illustrated in Table 1. Compared with the V-disparity [11] and sub-V-disparity [12] approaches, the false positive rate (FPR) of our approach decreases by 1.79% and 1.85% respectively, the other indices of our approach get an improvement with an average of 7.47% and 3.30% respectively.

Both the V-disparity and sub-V-disparity approaches are based on the hypothesis that the ground pixels occupy the main part in each image row, which is not suited for the distant ground region. The V-disparity algorithm models the ground plane with fixed disparity value in each row, which fails in the situation with varying latitudinal slope. The sub-V-disparity algorithm alleviates the influence of the latitudinal slope by a sliding windows paradigm. However, the sliding windows paradigm is contradicted against the essence of V-disparity representation. The LDTD of our approach can extract the feature of the ground plane in various terrains without requiring the dominance of ground pixels.

3.4 Parametric Analysis

We evaluate the performance of our approach with different patch size and sub-block size. The experimental results can be seen in Fig. 6. The precision increases and the FPR decreases with the increase of the patch size, which is attributed to more contextual information being employed into the convolution neural network. The performance keeps stable when the patch size is pretty large with enough contextual information. As the sub-block size increases, the performance gets worse, which is attributed to the decreasing resolution of the Disparity Texture Map. In our experiment, the sub-block size is 1x1 and patch size is 96x96, which can achieve the best performance in precision and FPR.

4. Conclusion

We propose a novel approach for ground plane detection.

The approach proposes a new Local Disparity Texture Descriptor (LDTD) to extract the feature of ground plane directly from the disparity map. The contextual information is employed in the convolution neural network to improve the classification accuracy. Based on the above methods, our approach can address well most of urban terrains. The improved performance of our approach is demonstrated in the KITTI benchmark compared with the traditional methods.

Acknowledgments

This project was supported by Shanghai Sailing Program (No.17YF1427300).

References

- [1] X. Chen, K. Kundu, Y. Zhu, A.G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," *Proc. Adv. Neural Inform. Process. Syst.*, pp.424–432, 2015.
- [2] C.G. Keller, M. Enzweiler, M. Rohrbach, D.F. Llorca, C. Schnorr, and D.M. Gavrilu, "The benefits of dense stereo for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol.12, no.4, pp.1096–1106, 2011.
- [3] L. Qu, K. Wang, L. Chen, Y. Gu, and X. Zhang, "Free space estimation on nonflat plane based on v-disparity," *IEEE Signal Process. Lett.*, vol.23, no.11, pp.1617–1621, 2016.
- [4] A. Wedel, H. Badino, C. Rabe, H. Loose, U. Franke, and D. Cremers, "B-spline modeling of road surfaces with an application to free-space estimation," *IEEE Trans. Intell. Transp. Syst.*, vol.10, no.4, pp.572–583, 2009.
- [5] A.D. Sappa, F. Dornaika, D. Ponsa, D. Gerónimo, and A. López, "An efficient approach to onboard stereo vision system pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol.9, no.3, pp.476–490, 2008.
- [6] K. Schauwecker and R. Klette, "A comparative study of two vertical road modelling techniques," *Proc. Asian Conf. Comput. Vis.*, pp.174–183, 2010.
- [7] J.K. Suhr and H.G. Jung, "Dense stereo-based robust vertical road profile estimation using Hough transform and dynamic programming," *IEEE Trans. Intell. Transp. Syst.*, vol.16, no.3, pp.1528–1536, 2015.
- [8] S. Nedeveschi, R. Danescu, D. Frentiu, T. Marita, F. Oniga, C. Poccol, T. Graf, and R. Schmidt, "High accuracy stereovision approach for obstacle detection on non-planar roads," *Proc. IEEE INES*, pp.211–216, 2004.
- [9] F. Oniga, S. Nedeveschi, M.M. Meinecke, and T.B. To, "Road surface and obstacle detection based on elevation maps from dense stereo," *2007 IEEE Intelligent Transportation Systems Conference*, pp.859–865, 2007.
- [10] J.K. Suhr and H.G. Jung, "Noise-resilient road surface and free space estimation using dense stereo," *Proc. IEEE Intell. Veh. Symp.*, pp.461–466, 2013.
- [11] R. Labayrade, D. Aubert, and J.-P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation," *Proc. 2002 IEEE Intell. Veh. Symp.*, pp.646–651, 2002.
- [12] Y. Dai, W. Wang, and Y. Kawamata, "Complex ground plane detection based on v-disparity map in off-road environment," *Proc. 2013 IEEE Intell. Vehicles Symp.*, pp.1137–1142, 2013.
- [13] C. Zhang, L. Wang, and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," *European Conference on Computer Vision*, pp.708–721, 2010.
- [14] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.34, no.11,

- pp.2274–2282, 2012.
- [15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol.86, no.11, pp.2278–2324, 1998.
 - [16] C.C.T. Mendes, V. Frémont, and D.F. Wolf, “Exploiting fully convolutional neural networks for fast road detection,” *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp.3174–3179, 2016.
 - [17] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The kitti vision benchmark suite,” *CVPR*, pp.3354–3361, 2012.
 - [18] J. Zbontar and Y. LeCun, “Stereo matching by training a convolutional neural network to compare image patches,” *Journal of Machine Learning Research*, vol.17, pp.1–32, 2016.
 - [19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *Proc. 22nd ACM International Conference on Multimedia*, pp.675–678, 2014.
-