Stixel-based Traffic Scene Representation Using U-disparity

Lei Chen¹, Yingping Huang¹, Xing Hu¹, Hongjian Wei¹ and Ming Liu¹

¹School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
²huangyingping@usst.edu.cn

Abstract. The Stixel World is a valid approach of interpreting 3D traffic scenes. This paper presents a novel and simple method to establish the Stixel World by exploiting the properties of U-disparity. Unlike the conventional stixel extraction method, the proposed method does not model the ground plane, but construct a U-disparity map and accordingly remove ground-related points and background-related points in virtue of U-disparity properties. The resulting U-disparity map is re-projected back to the disparity map so that the base and top-points of foreground obstacles can be determined. Experiments show that the proposed method can effectively build a Stixel World for a variety of scenarios and is not constrained by topography and object classes.

1. Introduction

Scene understanding is a fundamental task for on-board driver assistance systems, enabling a car to be aware of its driving environment and warning drivers of potential hazards. Cameras are commonly used in these applications. In particular, stereo vision is capable of a powerful 3-dimensional reconstruction and therefore greatly contributes to these systems. Scene understanding is to recognize all moving and stationary objects and therefore determine the free space where the vehicle may potentially drive in. The Stixel World as proposed in [1] is one of valid approaches of interpreting a traffic scene. By using a binocular camera setup, it provides a column-wise segmentation of obstacles named “stixel”. Each stixel is a rectangular stick defined by its 3D position relative to the camera and stands on the ground, having a certain height. The existence of a stixel excludes free space at this position, and the free space is a subset of the ground manifold. The Stixel World method has been proved to be a compact and efficient representation of 3D traffic scene with regards to reduced data size and efficient computation for latter processing. This paper extends the work in [1] and proposes a new and simple stixel estimation method by exploiting the properties of U-disparity.

2. Related work

The Stixel World is first proposed by Hern´an Badino et al [1] in 2009. They computed a stochastic occupancy grid from dense stereo map. Free space was computed from a polar representation of the occupancy grid, and accordingly the base-point of the obstacles was obtained. To obtain the height of the stixels, they firstly segmented foreground and background in disparity image according to predefined distance knowledge, and then employed a cost function which relied on the greyscale information to determine the boundary between foreground and background. Since then, a number of approaches [2-7] were undertaken to generate the Stixel World representation. In ref. [2], the original static stixel world was extended to a dynamic Stixel World, in which lateral and longitudinal motion was estimated for each stixel by using the 6D-Vision Kalman filter framework. This extension was
useful for detection of moving obstacles. The stixel estimation has been also enhanced by using pixel level semantic segmentation based on color cues and the geometric features of the scene [3]. The approach was further extended utilizing convolutional neural networks [4]. Ref. [5, 6] proposed an approach that a stixel world model is computed directly from the stereo images without computing an intermediate depth map. Instead of computing disparity, they computed a cost volume as the sum of absolute differences over the RGB color channels using a window for every pixel. Ref. [7] constructed stixel using a collinear trinocular vision system. They used three conjugate stereo images to measure the consistency of disparity values by means of the transitivity error in disparity space.

The Stixel World has been used in some applications and proved to be a compact representation of 3D scene. Enzweiler et al [8] utilized the Stixel World as an early focus-of-attention stage and matched the 3D stixels with prior knowledge about the object class of interest including 3D geometry and symmetry for stixel-based object recognition. Ref. [9] extended depth stixels into semantic stixels by leveraging a deep learning-based scene labeling approach that providing a class label for each pixel. Ref. [10] presented a Bayesian segmentation approach by using the Dynamic Stixel World for detection and tracking of moving objects in traffic scenes. The Stixel World representation has been further enhanced by utilizing the online color modelling for the road versus obstacle segmentation [11].

3. Building the stixel world

When the Stixel World is applied to describe a stereo scene, obstacles closest to the car in all directions are marked as obstacles of interest, i.e. stixels. The obstacles behind the stixels are considered as background. The free space is regions in the ground manifold without any obstacles, i.e. regions ahead of the ego-vehicle where the vehicle may potentially drive in. Free-space and stixel calculations are closely related to each other because the existence of a stixel excludes free space at this position. The tasks of building a Stixel World are to distinguish stixels from the ground surface and their background, and determine their height and distance.

3.1. U-disparity formation

We formulate the U-disparity map by exploiting the U-V-disparity representation as described in ref. [12]. The dense disparity map $\delta(u, v), 1 \leq u \leq N_{col}, 1 \leq v \leq N_{row},$ can be obtained from the stereo image pair. Let $H$ be the function of the image variable $\delta$ so that $\delta_u = H(\delta).$ We call $\delta_u(u, d)$ as the "U-disparity" image. $H$ accumulates the points with the same disparity that occur on a given image column. For the image column $u,$ the grey scale of a point in $\delta_u(u, d)$ is the number of the points with the same disparity $d$ on the column, i.e.

$$\delta_u(u, d) = \text{card}\{u: 1 \leq u \leq N_{col} \cap \text{INT}(\delta(u, v)) = d\}$$

(1)

where $0 \leq d \leq d_{max}$ defines the quantized disparity range in the $N_{col} \times N_{row}$ disparity map $\delta(u, v);$ INT is the integer function. The abscissa of $\delta_u(u, d)$ is $u$ as the same as $\delta(u, v).$ The ordinates of $\delta_u(u, d)$ is the disparity $d.$

In a driving environment, objects can be divided into two categories. The first category is the obstacles vertical to the ground plane, including vehicles, pedestrians, trees and roadside constructions, which can be abstracted as vertical planes or vertical oblique planes (side planes). The second is road surface, which can be abstracted as a horizontal plane. These planes have different representations in the U-disparity map [13]. Vertical planes will appear as horizontal line segments and vertical oblique planes will appear as approximate oblique line segments in the U-disparity map. Because a horizontal plane covers a close-far region, the number of accumulated points with the same disparity in each column for a horizontal plane is dispersed, i.e. $\delta_u(u, d)$ is tiny. Thus, a horizontal plane will dispersedly distribute in the U-disparity map.

Fig. 1(left) is the left image of a pair of stereo images of a typical traffic scene. In this work, we utilize the stereo matching algorithm to generate a dense disparity map. The result is shown in Fig. 1(middle). The formulated U-disparity map is shown in Fig. 1(right), where the 3D obstacles are assembled in the horizontal or oblique line segments while the road surface is dispersed in the map.
3.2. Elimination of ground and background

Since the points of ground manifold have a small greyscale and will dispersely distribute in the U-disparity map, they can be removed by setting a suitable greyscale threshold. Fig. 2(a) shows the binarized U-disparity map with the ground-related points removed. Comparing Fig. 2(a) with Fig. 1(right), it can be observed that only the obstacle information is preserved as horizontal and oblique line segments. It is worth mentioning that this step also removes some discretized points with a deviated disparity, thereby making the distance estimation more accurate.

The position of a point cluster in Fig. 2(a) reflects its distance to the cameras, the lower the closer. To remove the background obstacles, we can retain the lower points and discard higher points on the same column in Fig. 2(a). However, in real traffic scenes, obstacles may contain areas lacking of texture, resulting in missing disparities in the dense disparity map. Consequently, the line segments corresponding to obstacles in Fig. 2(a) may discontinue. Therefore, we conduct the following conversion to yield a condensed U-disparity map.

Fig 2(a) is condensed as shown in Fig. 2(b). The condensed U-disparity map can be written as

\[ T(a, b) = \begin{cases} 0, & \text{if } \sum_{i=0}^{w-1} P(a \ast w + i, b) = 0 \\ 1, & \text{otherwise} \end{cases} \]  

(2)

where \( a \) and \( b \) respectively denote the abscissa and the ordinate of \( T \). \( P \) is the binarized U-disparity map, i.e. Fig. 2(a). The number of rows of \( T \) is the same as the one of \( P \), and the number of columns of \( T \) is \( 1/w \) of the one of \( P \). In Fig. 2(b), the position relationship between objects remains unchanged, but the points within the same objects become continuous.

And then, the points in each column of Fig 2(b) are traversed from bottom to top. The points with the earliest continuous nonzero values are considered as foreground objects and retained, and the points above are discarded as background-related points. The result is shown in Fig. 2(c).

The condensed U-disparity map with background removed is reconverted into a U-disparity map:

\[ Q(s, t) = \begin{cases} P(s, t), & \text{if } T \left( \text{INT} \left( \frac{s}{w}, t \right) \right) = 1 \\ 0, & \text{otherwise} \end{cases} \]  

(3)

where \( Q \) is the regenerated U-disparity map, \( T \) is the condensed U-disparity map with background removed, i.e. Fig. 2(c). The regenerated U-disparity map is presented in Fig. 2(d).

(a)                                         (b)                 (c)                                          (d)

Figure 2. Formation of resulting U-disparity map with background and ground removed.

3.3. Elimination of ground and background

The U-disparity map in Fig. 2(d) contains foreground obstacles with other points removed. The base and top-points of foreground obstacles can be extracted by projecting Fig. 2(d) to the disparity map.
A point in a U-disparity map is calculated from one or more points in its corresponding disparity map. Conversely, a point in the disparity map corresponds to a single point in the U-disparity map. This mapping relationship can be described in Fig. 3 as a 3D to 3D space mapping. In Fig. 3, $I_1$ is a 3D U-disparity map while $I_2$ is the 3D disparity map which $I_1$ is generated from. $I_1$ and $I_2$ have the same abscissa ($U$). The vertical coordinate ($D$) in $I_1$ is disparity value while the vertical coordinate ($V$) in $I_2$ is image row number. The third coordinate ($N$) in $I_1$ is the number of points with the same disparity in a column while the third coordinate ($D$) in $I_2$ is the same as the vertical coordinate in $I_1$. The abscissa and vertical coordinate define the position while the third coordinate represents the greyscale of the image. The coordinate system ($U$, $D$, $N$) in $I_1$ can be described as the following function:

$$N = \Phi(U, D)$$ (4)

The coordinate system ($U$, $V$, $D$) in $I_2$ can be represented by:

$$D = \Delta(U, V)$$ (5)

A point $P^*(a, b, c)$ in $I_1$ has multiple corresponding contribution points $P^*_1(i_1, j_1, k_1), P^*_2(i_2, j_2, k_2), \ldots, P^*_s(i_s, j_s, k_s)$ in $I_2$. We have the following relations:

$$i_1 = i_2 = \cdots = i_s = a$$

$$\Phi(a, \Delta(a, j)) = c \neq 0$$

$$k_1 = k_2 = \cdots = k_s = b$$

$$s = c$$

The set $j = \{j_1, j_2, \ldots, j_s\}$ can be determined by searching for the points with disparity $b$ on the $a^{th}$ column in $I_2$. Thus, the contribution points $P^*_1(a, j_1, b), P^*_2(a, j_2, b), \ldots, P^*_s(a, j_s, b)$ can be obtained. Points $P_1(a, j_1), P_2(a, j_2), \ldots, P_c(a, j_c)$ in the disparity map can be determined also.

Mapping Fig. 2(d) back to the dense disparity generates Fig. 4, where the ground-related points and the background-related points have all been removed with only foreground obstacles reserved. The extraction of base-points and background boundary are straightforward. As shown in Fig. 5, the base and top-points of foreground obstacles are marked in the original scene image.

3.4. Stixel extraction and distance estimation

Once the base and top-point for every column have been computed, the extraction of the stixel is straightforward. Given $m$ as the width of the image, all the base-points $\{b_1, b_2, \ldots, b_m\}$ and the top-points $\{t_1, t_2, \ldots, t_m\}$ are represented in Fig. 5. There are $\frac{m}{w}$ stixels in the image, where $w$ is the width of a stixel. For $i^{th}$ stixel, it contains base-points $B_i = \{b_{(i-1)w+1}, b_{(i-1)w+2}, \ldots, b_{(i-1)w+w}\}$ and top-points $T_i = \{t_{(i-1)w+1}, t_{(i-1)w+2}, \ldots, t_{(i-1)w+w}\}$. The stixel is described by a rectangle spanning from $[(i-1)w+1]^{th}$ column to $[(i-1)w+w]^{th}$ column and $\text{Min}(T_i)^{th}$ row to $\text{Max}(B_i)^{th}$ row. Taking $w = 10$, the Stixel World of the scene is established as shown in Fig. 6(left).

The distance of a stixel is estimated as the dominant distance of all contribution points within it. Since the deviation points have been removed during the processing of building stixels, each contribution point has an accurate disparity value. The greyscale information of the stixels in Fig. 5 encodes the change of the distances. The corresponding free space can be drawn as in Fig. 6(right).
4. Experiment
The algorithm runs in a PC equipped with a 2.4-GHZ Intel Dual Core i5 processor and 8GB of RAM. Fig. 7 displays the stixel representation and corresponding free space generated from the described method for city, residential, road and campus scenarios, which are selected from the KITTI benchmark database [14]. The image resolution is 1242*375 pixels. The greyscale information of the stixels encodes the change of the distance to the equipped car. It can be seen that the proposed algorithm is capable of identifying a variety of foreground obstacles and generate accurate stixel world for the scenarios. It is worth to mention that the proposed algorithm is not constrained on a flat road. Fig. 7(b) shows the stixel extraction result in a mild uphill pavement, and Fig. 7(c) shows the result in a potholes pavement. It can be seen the proposed algorithm can work well in those scenarios.

This work focuses on extraction of accurate stixel boundary rather than distance measurement. Therefore, we evaluate the algorithm with regards to the position error of the extracted stixels by comparing the detected base and top-points with the manually marked ground truth. The errors for each frame are computed as follows;

$$\text{RMSE}_1 = \sqrt{\frac{\sum_{i=1}^{N_{stx}} (B_i - b_i)^2}{N_{stx}}}$$
$$\text{RMSE}_2 = \sqrt{\frac{\sum_{i=1}^{N_{stx}} (T_i - t_i)^2}{N_{stx}}}$$

where $N_{stx}$ is the number of stixels in the image. $B_i$ and $T_i$ are the detected base-point and top-point position, respectively. $b_i$ and $t_i$ are the ground truth position. A sequence of image with 100 frames for each scenario is evaluated, and TABLE 1 shows the average errors.

5. Conclusion
This paper proposes a stixel estimation method by exploiting the properties of U-disparity. In a U-disparity map, obstacle-related points are assembled as horizontal or oblique line segments while ground-related points dispersedly distribute. By applying a greyscale threshold, the ground points can be eliminated. Background obstacles can be also removed by means of distance information. The resulting U-disparity map only contains foreground obstacles, and is re-projected back to the disparity
map so that the base and top-points can be determined. Once the base and top-point for every column have been computed, the extraction of the stixel is straightforward. The experimental results show that the proposed method can effectively build a Stixel World for a variety of scenarios.

Comparing with the existing Stixel estimation algorithms, the contributions of this work can be found as follows: 1) Instead of using a predefined distance threshold and greyscale-based cost function to segment foreground and background [1], the proposed method removes background and ground-related points in the U-disparity map and extracts the base and top-points of every column by re-projecting the resulting U-disparity map back to the disparity map. 2) The proposed method eliminates the need of modeling ground plane, thus, it is not constrained by topography. The algorithm can work well no matter on an incline road surface or a potholes pavement. 3) The process of removing ground-related points and compression (modularization) of the data in the condensed U-disparity benefit to the elimination of points with deviated disparities.

References
[1] H. Badino, U. Franke, and D. Pfeiffer, “The stixel world - a compact medium level representation of the 3d-world,” in DAGM Symposium, (Jena, Germany), September 2009.
[2] D. Pfeiffer and U. Franke, “Efficient representation of traffic scenes by means of dynamic stixels,” in IEEE Intelligent Vehicles Symposium, Proceedings, 2010, vol. 43, pp. 217-224.
[3] T. Scharwachter, and U. Franke, “Low-level fusion of color, texture and depth for robust road scene understanding,” in IEEE Intelligent Vehicles Symposium, Proceedings, 2015, vol. 2015–Augus, no. IV, pp. 599–604.
[4] D. Levi, N. Garnett, and E. Fetaya, “StixelNet: A Deep Convolutional Network for Obstacle Detection and Road Segmentation,” in British Machine Vision Conference (BMVC), Proceedings, 2015, pp. 1–12.
[5] R. Benenson, R. Timofte, and L. Van Gool, “Stixels estimation without depth map computation,” in Proc. IEEE Int. Conf. Comput. Vision. Barcelona, Spain, November 2011, pp. 2010–2017.
[6] Z. Wieszok, N. Aouf, O. Kechagias-Stamatis, and L. Chermak, “Stixel based scene understanding for autonomous vehicles,” 2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC), Calabria, 2017, pp. 43-48.
[7] N. H. Saleem, H. J. Chien, Mahdi Rezaei, and R. Klette, “Improved stixel estimation based on transitivity analysis in disparity space,” in Proc. 17th International Conf. Computer Analysis of Images and Patterns, August 2017, pp. 28-40.
[8] M. Enzweiler, M. Hummel, D. Pfeiffer, and U. Franke, "Efficient Stixel-based object recognition," in IEEE Intelligent Vehicles Symposium, Alcala de Henares, 2012, pp. 1066-1071.
[9] L. Schneider, M. Cordts, T. Rehfeld, D. Pfeiffer,and M. Enzweiler, “Semantic Stixels: Depth is Not Enough,” in IEEE Intelligent Vehicles Symposium, Gothenburg, Sweden, June 2016, pp.110-117.
[10] F. Erbs, B. Schwarz, and U. Franke , “From Stixels to Objects - A Conditional Random Field based Approach,” in IEEE Intelligent Vehicles Symposium, Gold Coast, Australia, June 2013, pp.586-591.
[11] W. P. Sanberg, G. Dubbelman, and P. H. N. de With, “Extending the Stixel world with online self-supervised color modeling for road-versus-obstacle segmentation,” in 17th IEEE International Conference on Intelligent Transportation Systems, Qingdao, 2014, pp. 1400–1407.
[12] R. Labayrade, D. Aubert, and J. P. Tarel, “Real time obstacle detection in stereovision on non-flat road geometry through "v-disparity" representation,” in IEEE Intelligent Vehicles Symposium, 2002, pp. 646-651.
[13] Z. Hu, F. Lamosa and K. Uchimura, “A complete U-V-disparity study for stereovision based 3D driving environment analysis,” in 5th International Conference on 3-D Digital Imaging and Modeling, Proceedings, 2005, pp. 204-211.
[14] A. Geiger, P. Lenz and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012, pp. 3354-3361.