Towards Scene Understanding Implementing the Stixel World

Amélie Grenier*‡, Alaa AlZoubi†, Luke Feetham* and David Nam*

*Centre for Electronic Warfare, Information and Cyber
Cranfield University, Defence Academy of the UK
Shrivenham, United Kingdom
†School of Computing
The University of Buckingham
Buckingham, United Kingdom

‡Corresponding author: a.grenier@cranfield.ac.uk

***Abstract*—In this paper, we present our work towards scene understanding based on modeling the scene prior to understanding its content. We describe the environment representation model used, the Stixel World, and its benefits for compact scene representation. We show our preliminary results of its application in a diverse environment and the limitations reached in our experiments using imaging systems. We argue that this method has been developed in an ideal scenario and does not generalise well to uncommon changes in the environment. We also found that this method is sensitive to the quality of the stereo rectification and the calibration of the optics, among other parameters, which makes it time-consuming and delicate to prepare in real-time applications. We think that pixel-wise semantic segmentation techniques can address some of the shortcomings of the concept presented in a theoretical discussion.**

***Index Terms*—Stixel, stereo, calibration, scene understanding, semantic segmentation.**

I. Introduction

Autonomous driving, for any type of vehicle, is a well established notion and has guided research objectives in the past decade. Combined with progress in the supporting technologies, it led to a rapid expansion in the development of self-driving cars and driving assistance systems, especially in the last couple of years. Manufacturers have announced their releases for the next decade.

To improve the autonomy of cars, we aim to provide them with a good level of intelligence in terms of analysing and understanding their surroundings. This intelligence mainly involves the capacity of automatic and autonomous interpretation of scenes, both geometric and semantic, including both static and dynamic objects. Multiple sensors of different modalities, such as cameras, radars and inertial navigation systems can be used. Exploiting these exteroceptive and proprioceptive sensors in analysing the scene and internal vehicle data to improve vehicle situation awareness and decision-making in all conditions is crucial.

In our research we focus on the information we can retrieve from a pair of RGB cameras. In visual scene analysis, rough scene segmentation is usually done by first dividing the image into sub-regions then defining appropriate features for the desired classes of objects. This implies that the objects are detected before the feature extraction is performed to classify them. It implicitly means that semantic understanding is based on geometric understanding. Earlier, these two interpretations were separate and successive problems. For example, a well used method for object representation is the Stixel World [1], principally chosen for its robustness, compactness and precision in depicting the content it detected in stereo-images. After the return of deep learning in 2012, semantic information was consecutively added to the geometric model [2] [3]. In the last few years, many object detection approaches were developed combining the object recognition aspect with the sub-regions definition, and the convolutional neural network features extractor; for instance the R-CNN [4] to its recent evolution, the Mask R-CNN [5]. With deconvolutional neural networks [6], pixel-wise semantic segmentation [7] can be used for object recognition, without needing a classic object detector. Nowadays, it is common to encounter the fusion of both detection and segmentation problems [8] [9], including using the Stixel World approach. However, its usage seems to reduce to the advantage of deep learning techniques.

Following this trend, our research objective is to combine detection and recognition to provide scene understanding for self-driving cars. In order to do so, we implement, on a car, a recent version of the Stixel approach - the work of Wieszok *et al.* [10]. We present our work to adapt it and the issues we encountered. We discuss reorienting our research towards pixel-wise semantic segmentation using deep learning instead of the Stixel World, despite evident and appropriate qualities for environment representation in autonomous vehicles applications.

The rest of the paper is divided as follows: we present our implementation's choices, our experiment and the algorithm used in section II. We then show our preliminary results and discuss the techniques used in section III, to conclude in section IV.

II. Case Study

The work of Wieszok *et al.* [10], an extension to the original work of Benenson *et al.* [11], used rectified pairs of images from the KITTI dataset [12] along with the corresponding calibration parameters. To adapt it to real-time image acquisition, we made equipment choices and merged the existent method with the image acquisition and stereo-images rectification steps.

A. The Stixel World - algorithm

The Stixel World, originally presented by Badino *et al.* [1], aims to represent the free-space ahead of the vehicle in the form of sticks. The assumption is made that each obstacle can be approximated by vertical surfaces above the ground. Each stick, with its height and its distance from the ego-vehicle, is called a "stixel" (Fig.1). This method [1] computes stixels from processing the dense disparity map obtained from Semi-Global Matching [13]. As depth maps computation was time and resource consuming, it was replaced by Benenson *et al.* [11] by a matching cost volume (MCV) computation coupled with processing of the V-disparity image, introduced by Labayrade [14]. This results in very similar performances and a faster response time.

We apply a recent adaptation by Wieszok *et al.* [10] in our work, as depicted in Fig. 1. The algorithm computes the MCV as the absolute differences over the colour channels between the left and right rectified images, for every pixel and for every disparity value possible. Therefore, a low cost region should be observed at the correct disparity.

The first step retrieves the horizon line and estimates the ground plane by projecting MCV to the V-disparity image. The probability of the pixel belonging to the road and an Online Learned Colour Model (OLCM) [15] increases the contribution of the road in the V-disparity computation, and Iteratively Reweighted Least Square (IRLS) is used for the line estimation [10].

The second step estimates the base of the stixels, i.e. the distance to the nearest obstacle and the limit of the free-space on the ground plane previously computed. It projects MCV to the U-disparity domain and finds the low values in the stixel cost (Fig. 1d). The latter also uses the road probability and the OLCM.

The third and final step is the stixels' height estimation, based on the likelihood that each pixel above the ground belongs to the same estimated stixel's disparity. This step includes a colour membership function [10].

B. Set-Up, Implementation and Experiment

For these tests, we chose the equipment for a stereo-vision system allowing us to focus on the distance range of interest: from 2 to 50 m ahead. The specifications are:

- RGB cameras: CMOS sensor, 2448 x 2048 resolution, 38 FPS, 3.45 μm x 3.45 μm, global shutter, external trigger available, C/CS-mount compatibility, 2/3-in format.
- Lenses: focal length 8 mm, F 2.4-16, locked focus, 2/3-in format, C-mount.

The stereo set-up was decided considering the overlap required for the algorithm used and the advice from Gallup *et al.* [16]; we experimented with baselines from 40 cm to 55 cm. We use a software trigger for the image acquisition and perform the calibration using the AMCC toolbox provided by Michael Warren [17]. We implement the stereo-rectification from OpenCV [18], due to its capability of providing real-time performance, which is necessary for the application.

Equipped with the imaging system depicted above, we drove over a couple of miles taking a set of 271 stereo images. A calibration of the equipment was performed for each experiment. The environment is a set of flat parking lots and slightly sloped roads. The objects contained in the images are common for a driving scene: pedestrians, vehicles, poles, road signs, buildings, and speed breakers.

III. Results and Discussion

A. Analysis of the Stixel approach

The Stixel World was primarily chosen for its superiority in some aspects required for autonomous vehicle navigation [1]. The necessary obstacles' information - distance, location and height - is completely retained in a compact format. The obstacles are also detected and estimated precisely within the first 10 m of distance, between 0.5 and 1 m of error [10] [19].

However, error tends to increase beyond 20 m. Moreover, the concept comes with two shortcomings. The first is that the detected obstacles are only the first ones surrounding the ego-vehicle, as the primary objective of the original method is to delimit the free-space. Therefore, as accurate as the detection can get, it is not exhaustive. The second is that the stixels' width is predefined to an arbitrary value, usually 5 pixels for a trade-off between precision and computation time. It means that the stixels do not adapt to the boundary of the objects on the horizontal axis. While efforts have been made to address the former [20], to our knowledge none has been made to address the latter.

Another strong characteristic of the original Stixel World is its robustness as their method was shown as "quite parameter-insensitive" [20]. Yet, our results using the modified method [10] show that the addition of the numerous parameters brought instability. The performance of each step is dependent on the previous one (Fig. 1), which is not desirable as it increases the probability of failure. We also show that it cannot be applied in every scenario due to the conditions it requires as discussed below.

The surrounding environment is assumed to be planar as the horizon line is projected on a single line of the original image. As shown in Fig. 2c, it leads to incorrect estimation: on the left side of the image, the horizon was estimated under the base of the vehicles, and was estimated over the horizon on the right side. In Fig. 2d, the road is sloping downwards, the obstacles are estimated higher than they actually are as the algorithm searches the base of the obstacles on a planar ground plane.

The stereo-vision system has to be parallel to the horizon. Any speed breaker creates a disruption in the obstacles'

Fig. 1. Pipeline of the used method in this paper. What we modified from [10] is in dashed boxes, and what we have kept is within continuous lines. I_L and I_R are respectively the intensity in left and right images, u a given column, v a given row and d a given disparity. The image a) is the left image with the estimated horizon line, in red, and ground plane, in yellow; b) is the corresponding V-disparity and line estimation, c) is the estimated stixels' bases; d) the corresponding stixel cost and e) is the stixels' height estimation. Yellow are close obstacles, red are far. The red arrows represent performance's influences.

estimation. In Fig. 2e, the effects are shown while going up. The sensitivity to **the calibration** was already mentioned in the evaluation of the stixels made by Pfeiffer *et al.* [19] that showed "issues concerning our calibration and stereo processes that deserve dedicated consideration". Further investigation would also be required to study the impact of the use of the software trigger. Nonetheless, with the MCV being computed line by line, even an error of two lines in the rectification, which is usually acceptable, adds imprecision and leads to incorrect ground plane estimation and by proxy, to a wrong stixels estimation.

In an ideal scenario, like in the KITTI dataset and in Fig. 2a and b, all these conditions are met. However, errors leading to missed obstacles can still occur (Fig. 1e) when the lowest value picked in the stixel cost (Fig. 1d) is incorrect. This is due to the smoothness term parameter, which prevents incoherent jumps in depth.

Excluding the above mentioned environmental requirements and cameras' calibration, the list of parameters include: the parameters of the IRLS approximation technique for the line estimation in step 1; the road probability mask and the OLCM; the smoothness terms in steps 2 and 3 to balance the object and ground cost, and the membership function respectively; the scale factor; the colour membership function; the default height estimation; etc.

In addition, the computation time was 25 Hz in the work of Benenson *et al.* [11] against 15 Hz for the original method [1]. The method used here reduces the algorithm's speed. For example, the line estimation in the V-disparity went from 5.9 ms to 17.6 ms in one case. Due to the height of the images, the height estimation alone took 37.8 seconds on average, which is not appropriate for real-time applications. The reason why this method did not use a disparity map was for the purpose of speed. This is now nullified and the results do not justify its usage.

Indeed, in the majority of our images, the stixels estimation led to missed obstacles. Considering the conditions to the stixels computation and all the parameters, we believe this version of the Stixel World lacks stability and robustness and that the algorithm should generalise to more environment types. A self-driving car application requires higher detection rate and a more flexible approach.

B. Using Semantic Segmentation

The main objective of this research was to have an awareness of the surroundings using stixels, which included a geometrical and semantic understanding of the scene. However, with recent deep learning techniques, it is not necessary to have a geometric model of the objects to understand and recognise it, sequentially. Subsequently, we tested the semantic segmentation network called SegNet [7] on our dataset. Due to computation restrictions, we trained the network on a binary classification task - a vehicle detector - rather than on the original 11 classes. We trained the network on the KITTI dataset [12], the CamVid dataset [21] and the Cityscapes dataset [22].

Although it required training, inference with SegNet is faster than the stixels computation time. Our preliminary results are shown in the bottom row of Fig. 2. All the vehicles are detected, which was not the case with the Stixel approach. The detection is more exhaustive and the partially occluded objects are detected, not only the nearest ones. The network offers more flexibility: changes in the scene content have less impact on the results, making it more stable and robust in more diverse and uncommon driving scenarios. However, pixel-wise semantic segmentation does not compress the information like in the Stixel approach.

With SegNet, thanks to a pixel-wise segmentation, the objects are separated at the pixel level semantically and is object specific. That is not possible with the stixels as their width is fixed and retrieving boundaries would require post-processing. Although, whether Stixel World or SegNet is used, post-processing is necessary to retrieve instance information. Obviously, the two techniques are made for different purposes. So distance information is included in the Stixel approach but not with SegNet. Inversely, the object classification is covered with SegNet but not with the stixels computation on its own.

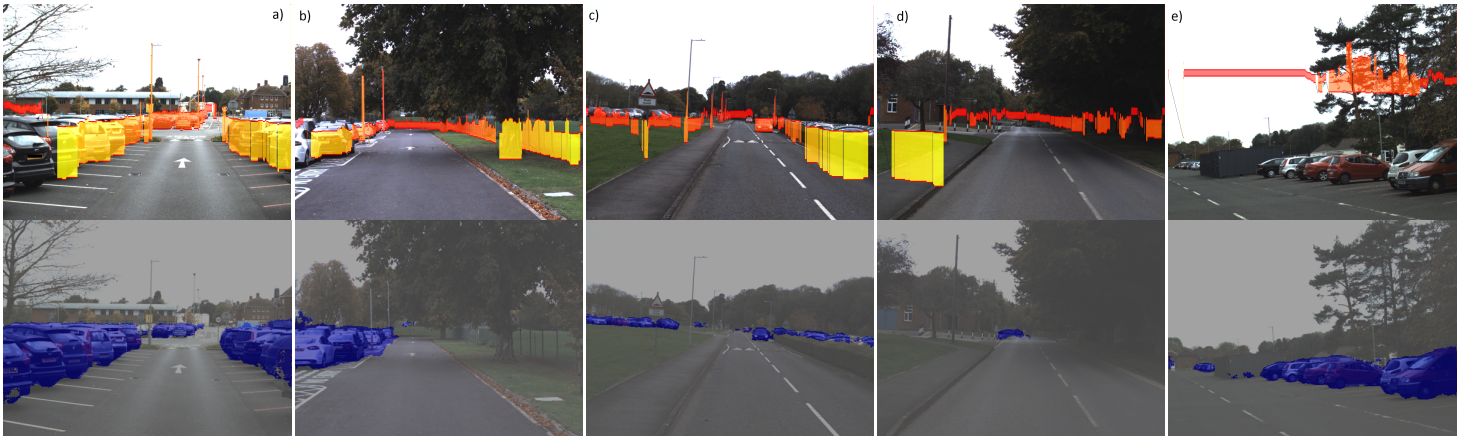

Fig. 2. Examples in applying the Stixel World (top row) and a comparison with semantic Segmentation - SegNet (bottom row). Cases of success of the Stixel approach are a) and b) while c), d) and e) are cases of failure. All the images come from the same experiment, with the same parameter settings. Error on the left of the images comes from the computation of the MCV, which creates a high cost zone in the Stixel Cost (Fig. 1d) on the left.

IV. Conclusion and Future Work

We presented an implementation of the Stixel World as a first step towards scene understanding. Its implementation comes with many parameters to consider, despite the appropriate theoretical environment representation aspects. In future work, using pixel-wise semantic segmentation, the objects could be semantically identified before using a grouping technique to detect and localise each instance.

Acknowledgment

A special thanks to Prof Nabil Aouf, Prof Mark A Richardson and Dr Lounis Chermak.

References

[1] H. Badino, U. Franke, and D. Pfeiffer, "The Stixel World - A Compact Medium Level Representation of the 3D-World," in *Pattern Recognition. DAGM 2009. Lecture Notes in Computer Science, vol 5748*. Springer, Berlin, Heidelberg, 2009, pp. 51–60.

[2] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, "Stixmantics: A medium-level model for real-time semantic scene understanding," *Lecture Notes in Computer Science*, vol. 8693 LNCS, pp. 533–548, 2014.

[3] "Semantic Stixels: Depth is not enough," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 110–117.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 580–587.

[5] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2980–2988.

[6] H. Noh, S. Hong, and B. Han, "Learning Deconvolution Network for Semantic Segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1520–1528.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous Detection and Segmentation," in *Computer Vision - ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8695*. Springer, Cham, 2014, pp. 297–312.

[9] L. Chen, Z. Yang, J. Ma, and Z. Luo, "Driving Scene Perception Network: Real-time Joint Detection, Depth Estimation and Semantic Segmentation," 2018.

[10] Z. Wieszok, N. Aouf, O. Kechagias-Stamatis, and L. Chermak, "Stixel Based Scene Understanding for Autonomous Vehicles," in *2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC)*. IEEE, 2017, pp. 43–48.

[11] R. Benenson, R. Timofte, and L. Van Gool, "Stixels estimation without depth map computation," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 2010–2017.

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[13] H. Hirschmüller, "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 807–814.

[14] R. Labayrade, D. Aubert, and J.-P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 2. IEEE, 2002, pp. 646–651.

[15] 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 *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, 2014, pp. 1400–1407.

[16] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, "Variable baseline/resolution stereo," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[17] M. Warren, D. McKinnon, and B. Upcroft, "Online calibration of stereo rigs for long-term autonomy," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, may 2013, pp. 3692–3698.

[18] A. Kaehler and G. Bradski, *Learning OpenCV 3, Computer Vision in C++ with the OpenCV Library*. O'Reilly Media, 2016.

[19] D. Pfeiffer, S. Morales, A. Barth, and U. Franke, "Ground truth evaluation of the Stixel representation using laser scanners," in *13th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2010, pp. 1091–1097.

[20] D. Pfeiffer and U. Franke, "Towards a Global Optimal Multi-Layer Stixel Representation of Dense 3D Data," in *Proceedings of the British Machine Vision Conference 2011*. British Machine Vision Association, 2011, pp. 51.1–51.12.

[21] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[22] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 3213–3223.