An environmental perception and navigational assistance system for visually impaired persons based on semantic stixels and sound interaction

Juan Wang, Kailun Yang, Weijian Hu and Kaiwei Wang
College of Optical Science and Engineering
Zhejiang University
Hangzhou, China
zjuwjopt@zju.edu.cn

Abstract—**Assistive technologies aim at enhancing personal mobility of individuals with disabilities to improve their independence and access to social life. For the visually impaired, perception during navigation comprises a major ingredient of independent living. With the development of computer vision, it is possible to meet the richer needs of visually impaired people. However, research on navigation assistance for the visually impaired is still relatively unexplored when compared with the active progress in autonomous driving which is already in full swing. In respond to this issue, we aim to leverage the study of the Stixel-World for automotive systems and transfer it to develop assistive technology for visually impaired people. The impressive research results of deep learning also suppose benefits for vision-based technology. Precisely, semantic segmentation is a task that enables identification of different objects uniformly. Inspired by these observations, we design a set of wearable visual aids, while the core algorithm is based on the stixel representations for three-dimensional world combined with pixel-wise semantic segmentation. Predetermined conditions for stixels in automotive research are optimized to fit the needs of navigation assistance in our algorithm, along with the incorporation of traversability-related semantic information. We also propose a sound mapping scheme, so that the environmental awareness about geographic and semantic information are conveyed to the visually impaired through acoustic feedback.**

Keywords—Stixel-World; semantic segmentation; sound mapping; wearable device

I. INTRODUCTION

As World Health Organization estimated, there are 253 million people living with vision impairment, while 36 million are blind and 217 million of them have moderate to severe reduced vision [1]. Vision is the most important source of human perception of the outside world. However, because of the limited ability of visually impaired people to access visual information, their independent navigation is greatly affected.

In recent years, a large part of researches are dedicated to the fields of machine vision. Visual aids based on camera input images emerge, but the market of vision-based navigation assistance for the visually impaired remains small. These devices usually acquire depth information, color images or infrared images through cameras, and distinguish the obstacles from the ground by the means of depth segmentation, edge detection or area growth, so as to guide the assistive navigation[2][3]. However, due to the inconsistent features and models, the single use of some kind of graphics processing method cannot provide comprehensive perception of the scene. What's more, the integration of the different threads is often computationally intensive, and it is difficult to ensure real-time assistance.

At the same time, autonomous driving attracts more attention with the emergence of a large number of valuable researches. Considering the similar needs for obstacle avoidance and scene understanding, the research for autonomous vehicles could to be leveraged to provide navigational assistance for the visually impaired. To overcome the limitation of incompatible assumptions across application domains, [4] clustered the normal vectors in the lower half of the field of view, [5] followed the Manhattan World stereo [5] to obtain ground-to-image transformation, while [6] integrated Inertial Measurement Unit (IMU) observations along with vision inputs in a straightforward way. In our research, semantic segmentation is used to realize the application of Stixel-World [8] in visual aids for the visually impaired, as shown in Fig.1.

Stixel-World [8][9][10] marked a significant milestone for flexibly representing traffic environments. Obstacles and free space are pixel-wisely segmented, and the stixel-level representation could provide compact and robust environmental awareness with respect to the original depth image. The stixel computation algorithm greatly reduces the computation of dense depth information and meets the speed requirements of time-critical applications. In addition, sonification of stixels are inherently smoother than representation of pixels for sound mapping. In this regard, it's

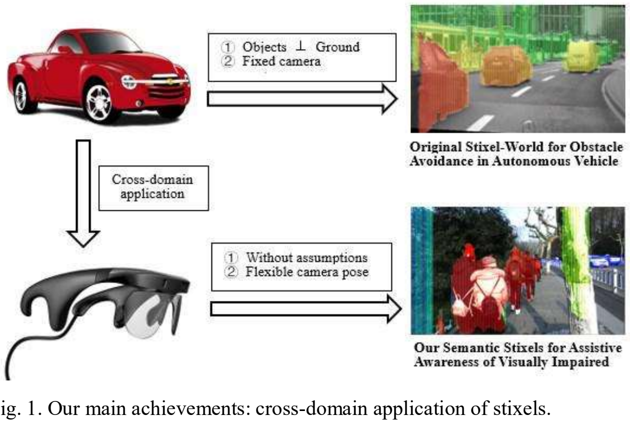
① Objects ⊥ Ground
② Fixed camera
Original Stixel-World for Obstacle Avoidance in Autonomous Vehicle
Cross-domain application
① Without assumptions
② Flexible camera pose
Our Semantic Stixels for Assistive Awareness of Visually Impaired

Fig. 1. Our main achievements: cross-domain application of stixels.

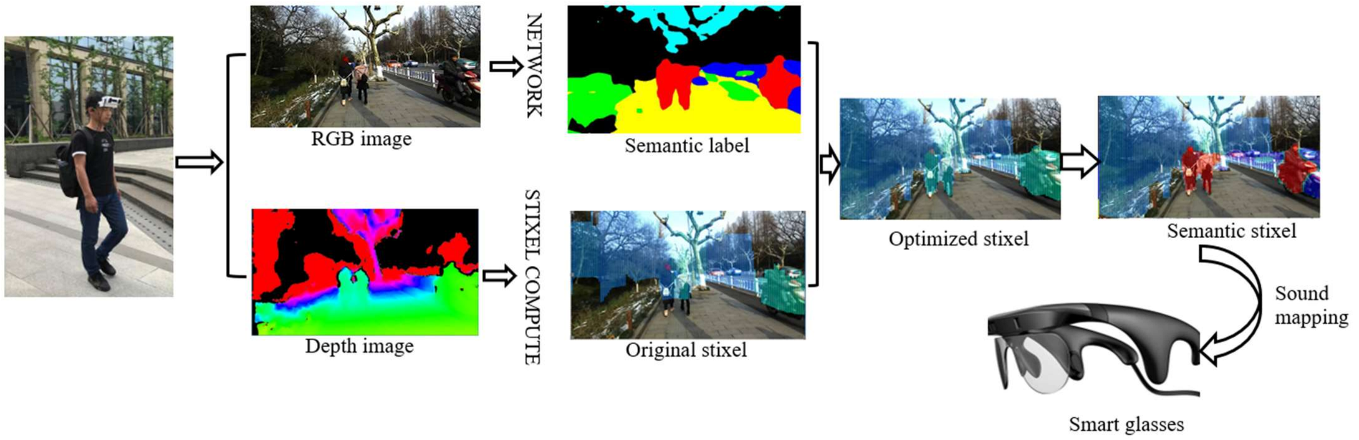

Fig. 2. The wearable navigation system as a pair of smart glasses consisted with a RGB-D camera, a PC in backpack and a pair of bone conduction headsets.

possible and beneficial to leverage the stixel-based techniques for autonomous vehicles and transfer them into assistive technology for the visually impaired.

Recent advances in deep learning have achieved breakthrough results in most vision-based tasks including semantic segmentation [10], which has grown as the key enabler to cover navigation-related perception tasks in a unified way. However, pixel-wise semantic segmentation was not usable in terms of speed. As presented in previous work [11][13], we have developed real-time semantic segmentation to assist visually impaired pedestrians.

In this paper, we extend previous established proof-of-concept to jointly infer the geometric and semantic layout of unknown scenes faced by visually impaired persons, which is similar to the semantic stixels scene model designed towards autonomous driving [13]. In order to convey the environmental information to the visually impaired, according to the study of human perception of sound information [15][15][16], we have designed a sound mapping scheme. In this contribution, visually impaired users are free to choose the assistance mode independently. One of the mode is the path guidance, which beeps the front obstacle to remind the visually impaired to avoid the hazardous obstacle, and this model does not contain semantic information; the other mode is environmental perception, while the images captured by the camera will be converted into sound, and the different categories in the scene are mapped using different musical tones. In both modes, distance and direction of obstacles are mapped by the combination of loudness tones and phase differences of the sound source.

The remainder of this paper is structured as follows. In section II, the navigation assistance framework is elaborated in terms of the wearable assistance system, the semantic segmentation, stixel architecture and the sound interactive design. Section III evaluated the approach and discussed the real-time and real-world performance of the system. In section IV, relevant conclusions are drawn and future works are expected.

II. System design

A. Wearable navigation device

In this work, we choose the ZED stereo camera [18] as the core device for vision data acquisition. The stereo camera's field of view is 0-110°, and we are able to acquire both depth and color maps. More importantly, the camera can obtain accurate object depth information from 0.5m to 20m, indoors and outdoors. In this regard, the RGB-D sensory awareness could meet our requirements for the field of view and the dynamic range.

How to convey the rich source of processed environmental information to the visually impaired? As is known to all, hearing is an important way for people to obtain information from the outside world, and it is believed in psychology that the information people receive from the outside world contains about 15% from the auditory channel. For this reason, auditory perception is a good visual aid to provide efficient feedback. Based on this knowledge, we use bone conduction headphones to convey the sound of image information mapping. This is important as visually impaired people need to continue hearing environmental sounds and the bone conducting interface allow them to hear a layer of augmented acoustic reality that is superimposed on the environmental sounds. Considering the portability and data acquisition needs, we design the device as head-mounted glasses [18] to acquire environment information and interact with visually impaired.

As is worn by the user in Fig. 2, the device is composed of a pair of smart glasses and a laptop in the backpack, and the smart glasses are contained with a stereo camera and a pair of bone conduction headphones. The camera captures real-time RGB-D streams while the RGB images are fed to the network for semantic segmentation and the depth images are used to compute stixels. The bone conducting earphones transfer the detection results for terrain awareness and collision avoidance. We utilize a laptop with Core i5-7200U processor and a cost-effective GPU 940MX as the computing platform, which could be easily carried in a backpack and is robust enough to operate in rough terrain.

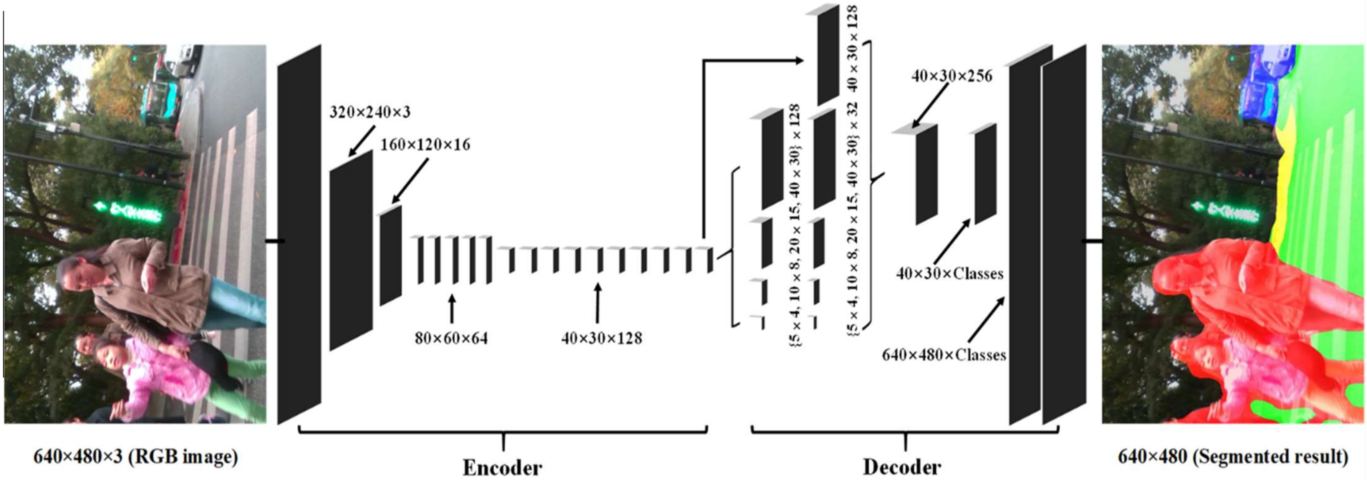

Fig. 3. Real-time semantic segmentation architecture we use, from left to right: (a) Input, (b) Encoder, (c) Decoder, (d) Prediction

B. *Semantic segmentation algorithm*

Visually impaired persons often rely on tactile sense to perceive different objects due to their imperfections in sight. However, during individual navigation, they need to make use of external aids to perceive the environment.

With the development of deep learning, a unified environmental awareness becomes possible. Semantic segmentation classifies a wide variety of scene classes, which directly leads to pixel-wise understanding. As mentioned in [11][13] we design the architecture according to the encoder-decoder architecture like SegNet, ENet and ERFNet [19][20][21] to respond to the surges in demand. Fig. 3 contains a depiction of the feature maps generated by each of the block in our architecture, from the RGB input to the pixel-level class probabilities and final prediction [11][13]. More specifically, our customized architecture is built on factorized convolution, sequential dilation and pyramid representation, which learns hierarchically high-level features that allow us to precisely detect semantic patterns and infer dense pixel-wise predictions, attaining the coverage of various scene elements. Based on the architecture tailored for the safety-critical context of blind assistance, the semantic segmentation algorithm can distinguish people, cars, sky, roads and sidewalks in a unified way, while maintaining the real-time speed with a good trade-off between efficiency and accuracy.

C. *Stixel algorithm*

The original Stixel-World simply separates obstacles from the freespace and background assuming that the surface of an obstacle is vertical and the baseline for all obstacles is on the ground. The position and tilting angle of the camera are fixed during autonomous driving, so the two important input parameters are set to constants. All of these factors limit the application of the stixel in the assistance of visually impaired persons.

However, the camera tilting angle changes constantly over time in visually impaired assistive devices. When people walk, obstacles are no longer vertical all time and some of the obstacles in the captured image is not connected with the ground. Under these circumstances, the original stixel algorithm generates false detections and missed detections of obstacles and passable areas. This problem represents a potential danger to the visually impaired user. What's more, stixel-based approach is insensitive to distant obstacles, and directly determines them as part of the background. However, for the visually impaired, some distant but fast-moving obstacles such as driving vehicles need to be perceived in advance.

Based on above analysis, we aim to improve the stixel algorithm by combining the masks of semantic segmentation to ensure the accuracy of the results and provided semantic awareness. At the same time, we directly calculate the traversable area, which greatly simplifies the calculation of free space computation [22][23]. Fig. 4 shows our optimization results of the stixel segmentation.

The specific algorithm is as follows. All input images are processed at a resolution of 640×360. Refer to the evaluation

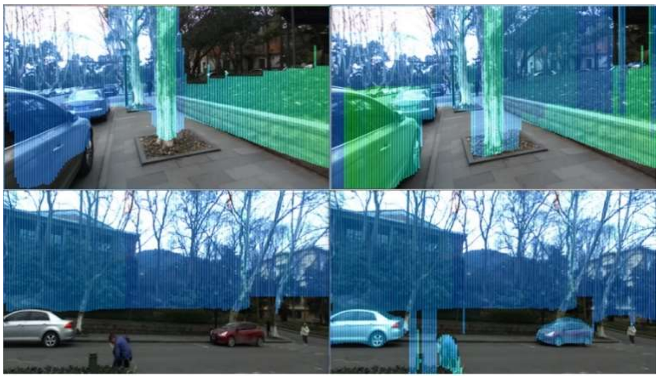

Fig.4. The left pictures are original stixels, and the right ones are our optimized stixels. In the above group, the car in the left figure was divided into the background, resulting in inaccurate depth information; in the bottom group, the distant vehicles and near incomplete people were not detected as obstacles. Our results solved these problems.

TABLE I SOUND MAPPING RULES OF AVOIDANCE MODE

Stixel	**Sound**
Depth/m	**Loudness**
>9	0
<3	1
3~9	Loudness decreases linearly with distance
Column coordinate	**Phase difference**
<160	-90
160~480	0
>480	90
Height	**Frequency**
-	1

of different stixel widths in [13], stixel width is set to 5 to strike a good balance between flexibility and efficiency.

Taking the baseline calculation of stixels as an example to introduce our algorithm. In order to extract stixels, a frame of image is scanned from the bottom up. According to the label made by semantic segmentation， if the column has a pixel of a car or a person, the row coordinates of the first point of the scanned vehicle or person are accumulated, and if there is not, the line will be scanned downwards and row coordinates of the first ground points are accumulated. Scanning five columns and the sum of row coordinates after averaging is the baseline of a stixel. The calculation of the topline follows the same principle. The depth of the stixels adopts the original calculation method, that is, the depth value of each stixel is a depth average value of all pixels included in the stixel.

At the same time, according to the results of the semantic masks, we distinguish the stixels between cars and pedestrians. Since the current semantic segmentation does not distinguish other types of obstacles, they are uniformly identified as general obstacles, as introduced in section III.

D. Sound mapping and interactive programs

As mentioned in section I, due to the abundant visual information acquired with our approach, we have designed two assistive modes, hazard avoidance and environmental awareness, accordingly, two sound mapping schemes are designed. After the device is turned on, the visually impaired hear the voice prompt “Please select auxiliary mode” and can choose different modes of assistance on their own initiative to avoid erroneous physiological discomfort and information comprehension over long periods of time.

We design a stixel-to-sound signal mapping scheme. Each sound source signal has three dimensions including phase difference, loudness and timbre, respectively, reflecting the position, distance and category of the stixel.

In obstacle avoidance mode, all stixels of obstacles within 9 meters are mapped to water droplets, of which the loudness and phase difference are determined by the distance and direction of the obstacle. We use the sound of water droplets for feedback because it is not annoying and maintains mellow when adjusting the mapping parameters. In context-aware mode, vehicles are mapped to horn sound and pedestrians are mapped to bell sound, and other obstacles correspond to water droplets, which can provide visually impaired people with richer environmental information. The timbre of these three music sources sound quite different. In this regard, it is easy to learn the feedback to distinguish different categories of objects. The loudness of the sound ranges from 0 to 1, and the left and right channels have a phase difference of -90~90°, where 0°means that the obstacle is in front of the obstacle. Precisely, specific sound mapping rules are given in the table1.

III. EXPERIMENTAL TEST AND CONDUCTED USER STUDY

We collected a real-world dataset by the ZED binocular camera from the City College, Nanshan Road and Yuquan Campus in Hangzhou, China. While wearing the camera, we captured the image at 5FPS so the test results effectively reflect the actual feasibility of the entire system.

In order to measure the computational performance of our algorithm, we processed 100 frame disparity images and RGB colors, and at the resolution of 640 ×360 on our processor which is introduced in Section II(A), the average total computation time of a single frame is 36ms, while the image acquisition and preprocessing from the smart glasses take 3ms, the time cost for the semantic segmentation is 13ms, and the time cost for calculating stixels is 20ms.

Fig. 5 shows the result of our algorithm processing the scene image. Different colors of stixels represent different distances, while stixel colors encode disparities from close (yellow) to far (blue); in the semantic stixels, blue denotes the vehicle and red represents the pedestrian.

In our experiments, we use three different metrics that are designed to assess the viability of our semantic stixel model. First is the target recognition rate. We classify target obstacles into pedestrians (P), vehicles (V), and general obstacles (O), and collect statistics on the detection rates within the different ranges of distances. Second is the stixel redundancy rate, and the calculation formula is the ratio of the number of stixels corresponding to the stixels identified as obstacles and the number of actual obstacles. The third metric is the accuracy of

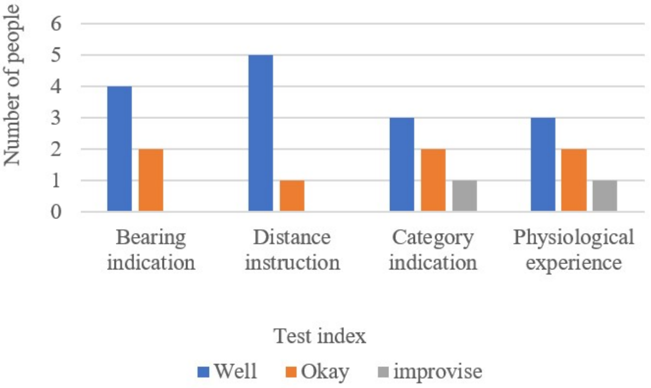

Fig. 6. User experience results.

the depth of stixels (DAR), which is the ratio of the stixels with an accurate boundary (the difference between the stixel boundary and the edge of the actual object within 10 pixels is considered accurate) in the stixels identified as an obstacle. We have randomly selected 100 frames of images for statistical purposes from the data sets of the three scenarios mentioned above. Test results are shown in the table2.

The experimental results show that the recognition rate of vehicles and pedestrians is higher, and the depth information is also more accurate even if the distance is far away. This is because the recognition rate is high and edge segmentation based on semantic segmentation are calculated well. It's beneficial for visually impaired people to avoid the danger of high-speed movement obstacles. For general obstacles, the upper boundary segmentation depends on the front segmentation algorithm in [9], so it is affected by the distance, the recognition rate is higher, and the depth calculation is more accurate within 3 meters. Although the recognition results from long distances are not satisfactory, it's sufficient for navigational assistance that the close static obstacles are prompted.

To evaluate the actual effect of the entire system and the performance of the sound mapping scheme, we also conducted a user experience test with six participants including three men and three women, participating in our test. They are blindfolded and perceived the environment based on the sound of our system.

We design the questionnaire about obstacle categories, azimuths, distances, and their physiological experience with sound effects. Each item was scored with "well", "okay" and "improvised", which means it needs to be improved. The test results are shown in the Fig. 6.

It turns out that in the obstacle avoidance mode, each person is able to discriminate the orientation of the obstacle (left, right or front) and the distance from the sound. In the context-aware mode, vehicles and pedestrians can be identified, but due to the superposition of multiple sounds, common obstacles are sometimes overlooked. After simple training, the participants can perceive more specific information of the obstacle distance based on loudness.

TABLE II SEMANTIC STIXEL ACCURACY

Category	Distance	Recognition rate	Redundancy rate	DAR
V	5~10m	91%	1.2	89%
	3~5m	90%	1	97%
	0~3m	95%	1	95%
P	5~10m	87%	0.9	96%
	3~5m	93%	1	90%
	0~3m	93%	1.1	91%
O	5~10m	80%	0.5	80%
	3~5m	85%	0.8	90%
	0~3m	90%	1.1	95%

IV. CONCLUSION AND DISCUSSION

In this work, we have designed a set of wearable visual aids that are mainly composed of a binocular camera (for image acquisition), bone conduction headphones (for voice interaction) and a computer in a backpack as a host of program processing. This set of equipment is of great help for the visually impaired, which enhances the mobility owing to the road guidance and environmental awareness.

Our image processing algorithms are based on stixels and semantic segmentation. The algorithm has been improved by combining the special needs of visually impaired persons, such as avoiding some inappropriate assumptions across application domains. In addition, we have designed a sound mapping scheme to convey the image information to the visually impaired.

At the same time, there is still room for improvement in our algorithm. The current approach relies on the results of semantic segmentation, which is a heavily researched topic where new datasets and network architectures will play an essential role. At present, there are some misjudgments in the semantic segmentation, such as mis-classifying roadblocks as pedestrians, although this has minor effect on obstacle avoidance. What's more, we currently have only solved a small part of the demand, next we will work to achieve the detection of small obstacles and other special scene like upstairs, to satisfy the needs of more complicated conditions such as airports and shopping centers.

In conclusion, it is worth promoting that the research of automatic driving technology is in full swing, while the research aimed to provide the visual aid for visually impaired people is so far relatively rare. Our research has proved the feasibility of application alliance and domain transfer. It can provide more assistance for the visually impaired by improving the technology originally designed for autonomous driving.

ACKNOWLEDGMENT

This work has been partially funded by the Zhejiang Provincial Public Fund through the project of visual assistance technology for the blind based on 3D terrain sensor (No. 2016C33136) and co-funded by State Key Laboratory of Modern Optical Instrumentation.

REFERENCES

[1] World Health Organization,"Vision impairmentand blindness," Available: http://www.who.int/mediacentre/factsheets/fs282/en/.

[2] K. Yang, K. Wang, W. Hu, J. Bai, "Expanding the detection of traversable area with RealSense for the visually impaired," Sensors, 2016.

[3] S. Wang, H. Pan, C. Zhang, and Y. Tian, "RGB-D image-based detection of stairs, pedestrian crosswalks and traffic signs," in Journal of Visual Communication and Image Representation, pp. 263-272, 2014.

[4] A. K. M. Roitberg, D.Stiefelhagen, "Using Technology Developed for Autonomous Cars to Help Navigate Blind People," in Computer Vision Workshop IEEE Computer Society, pp.1424-1432, 2017.

[5] H. C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarre, D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in Robotics and Automation (ICRA), pp. 6533-6540,2017.

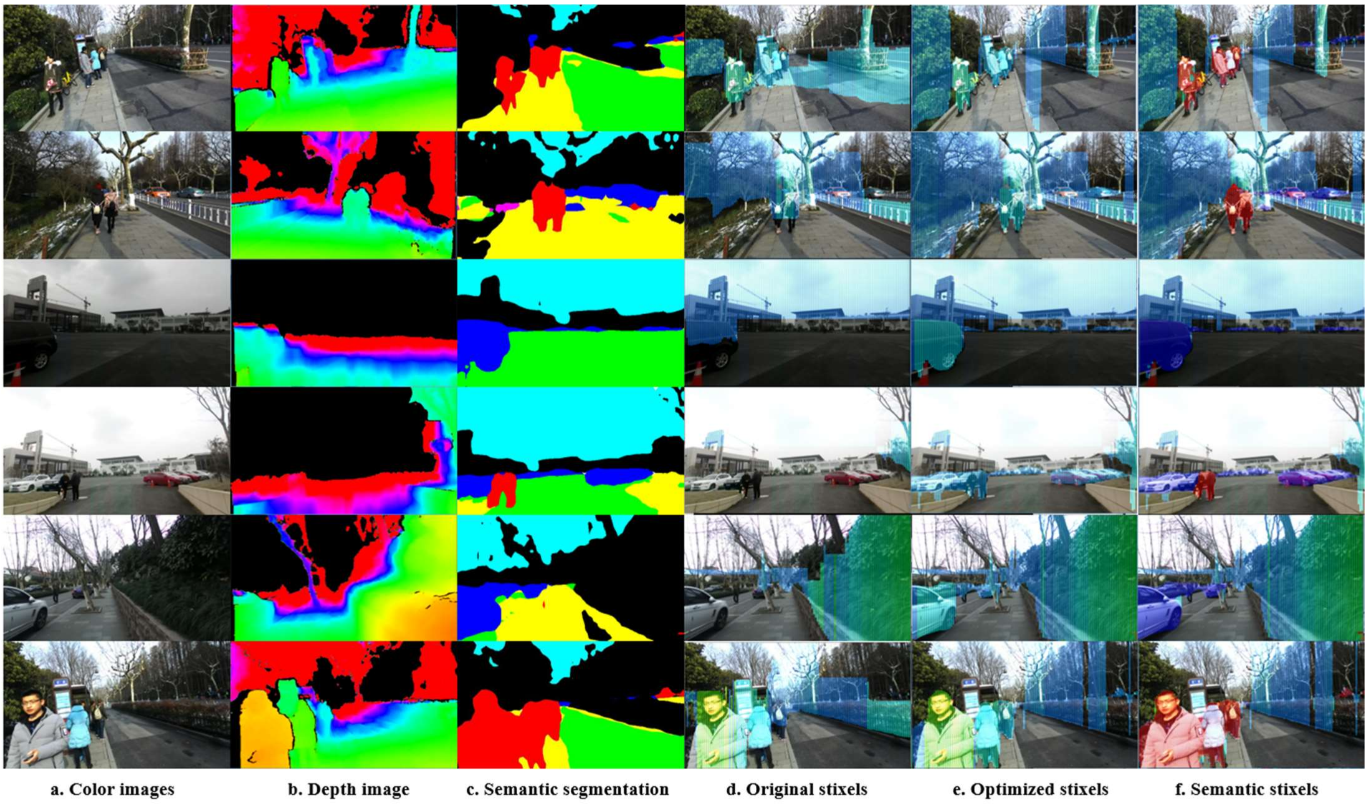

Fig. 5. Qualitative examples of the semantic stixels on real-world images produced by our approach compared with the original stixel.

[6] Y. Furukawa, B. Curless, S. M. Seitz, R. Szeliski, "Manhattan-world stereo," in Computer Vision and Pattern Recognition, pp. 1422-1429,2009

[7] K. Yang, K. Wang, R. Cheng, W. Hu, X. Huang, J. Bai, "Detecting traversable area and water hazards for the visually impaired with a pRGB-D sensor," in Sensors,2017.

[8] H. B. U. F. D. Pfeiffer, "The Stixel World-A Compact Medium Level Representation of the 3D-World," in Dagm Symposium on Pattern Recognition, pp. 5748 :51-60, 2010.

[9] U. F. D.Pfeiffer, "Efficient representation of traffic scenes by means of dynamic stixels, " in Intelligent Vehicles Symposium, pp. 217-224, 2010.

[10] B. Günyel, R. Benenson, R. Timofte, and L. Van Gool, "Stixels Motion Estimation without Optical Flow Computation," in Springer Berlin Heidelberg, pp. 528-539,2012.

[11] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," in International Journal of Automation and Computing, vol. 14, no. 2, pp. 119-135, 2017.

[12] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," In 2018 IEEE Intelligent Vehicles Symposium (IV), 2018.

[13] K. Yang, K. Wang, L. M. Bergasa, E. Romera, W. Hu, D. Sun, J. Sun, R. Cheng, T. Chen and E. López, "Unifying Terrain Awareness for the Visually Impaired through Real-Time Semantic Segmentation", Sensors, 18(5), p.1506, 2018.

[14] M. C. Lukas Schneider, Timo Rehfeld, David Pfeiffer, "Semantic Stixels: Depth is Not Enough," in Intelligent Vehicles Symposium IEEE, pp. 110-117, 2016.

[15] A. J. Kolarik, S. Cirstea, S. Pardhan, and B. C. Moore, "A summary of research investigating echolocation abilities of blind and sighted humans," in Hear Res, vol. 310, pp. 60-8, Apr 2014.

[16] N. Kopco, S. Huang, J. W. Belliveau, T. Raij, C. Tengshe, and J. Ahveninen, "Neuronal representations of distance in human auditory cortex, " in Proc Natl Acad Sci U S A, vol. 109, no. 27, pp. 11019-24, Jul 3 2012.

[17] S. R. Arnott, L. Thaler, J. L. Milne, D. Kish, and M. A. Goodale, "Shape-specific activation of occipital cortex in an early blind echolocation expert," in Neuropsychologia, vol. 51, no. 5, pp. 938-49, Apr 2013.

[18] Sterelabs, "Sense the world in 3D," Available: https://www.stereolabs.com/.

[19] N. Basoglu, A. E. Ok, and T. U. Daim, "What will it take to adopt smart glasses: A consumer choice-based review? " in Technology in Society, vol. 50, pp. 50-56, 2017.

[20] V. Badrinarayanan, A. Kendall and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation, " in IEEE Transactions on patter analysis and machine intelligence, vol. 39, no. 12, pp. 2481-2495, 2017.

[21] A.Paszke, A. Chaurasia, S. Kim and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation,"in arXiv preprint, 2016.

[22] E. Romera, J. Alvarez, L. M. Bergasa and R. Arroyo, "ERFNet:Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation," in IEEE Transactions on Intelligent Transportation Systems, 2017.

[23] H. Badino, R. Mester, T. Vaudrey, U. Franke, and A. G. Daimler, "Stereo-based Free Space Computation in Complex Traffic Scenarios," presented at the 2008 IEEE Southwest Symposium on Image Analysis and Interpretation, 2008.

[24] Oana.Ignat, "Disparity image segmentation for free-space detection," in International Conference on Intelligent Computer Communication and Processing IEEE, pp. 217-224, 2016.