Perception framework of water hazards beyond traversability for real-world navigation assistance systems

Kailun Yang[1], Luis M. Bergasa[2], Eduardo Romera[2], Juan Wang[1], Kaiwei Wang[1] and Elena López[2]

***Abstract*— Traversability perception constitutes an important task for robotics and visually impaired people, which aims to detect obstacle-free paths that allow individuals to ambulate with suitable navigation assistance. However, approaches that help prevent stepping into water areas are scarce in the state of the art. To address water hazard detection, this paper proposes a pRGB-D-SS perception framework, which incorporates: Polarization imaging (p), RGB-D sensory awareness and real-time Semantic Segmentation (SS). More specifically, as large water areas and small water puddles exhibit different characteristics, the detection of these two kinds of hazards pursue different pipelines. In our contribution, large water areas are detected together with traversable areas through pixel-wise semantic segmentation. Comparatively, the detection of water puddles extends the convolutional neural network based segmentation by using polarized RGB-D information as the primary cue. Beyond enhanced traversability awareness, it enables a unified framework of water hazard detection, which proves to be with qualified accuracy and speed for real-world assistance by experiments on two wearable systems including a pair of commercial smart glasses and a customized prototype.**

I. INTRODUCTION

Robotic vision have been widely leveraged to aid navigation in visually impaired individuals, creating a variety of personal assistive systems [1][2][3] to promote the awareness of traversability, which constitutes the backbone of mobile robotics [4] and blind assistance [5]. Beyond the proof-of-concepts established in these researches, the community has been motivated to provide more independence by integrating stairs detection [6] or crosswalk detection [7] at the basis of traversability analysis. In spite of the impressive strides towards higher mobility of the visually impaired, none of the approaches covers the perception of water areas that represent hazardous situations in everyday scenarios.

In the literature, a series of approaches addressed water hazard detection for robotics or self-driving vehicles [8][9][10]. Along with this research line, possibilities were investigated to leverage the developed techniques for robots [6] or autonomous cars [3][11], and transfer them into assistive technology for the visually impaired. However, almost all approaches produce intolerable side effects or rely on incompatible assumptions across application domains. For traversability awareness, underlying assumptions were frequently made such as the ground plane is the biggest part [1]. For water hazards detection, the ground was assumed to be horizontal and the incident light was assumed to be unpolarized in [9]. Moreover, variant versions of Manhattan World [3] or Stixel World assumptions [11] limit the flexibility in real-world applications. In the absence of these assumptions, water hazard detection is an ill-posed problem, forcing navigation systems into the trade-off between effectivity and accessibility, such as the method in [8], which requires mechanically rotating the linear polarizing filter and taking images at different angles.

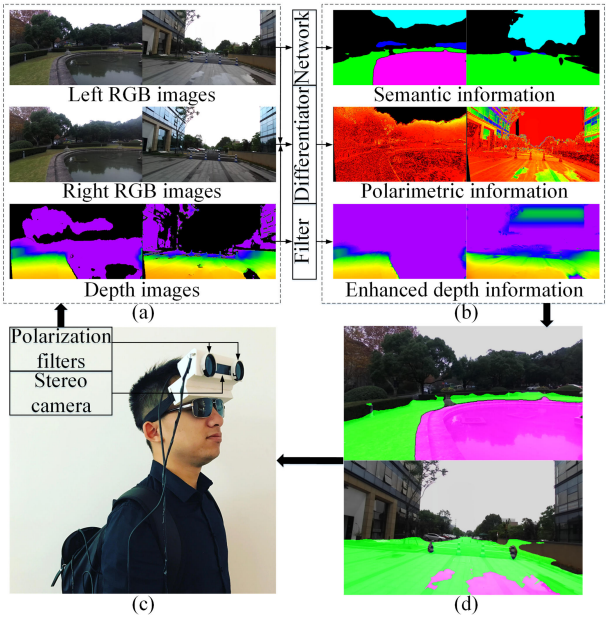


Fig. 1. Overview of the proposed pRGB-D-SS perception approach: (a) Original images from the sensor; (b) pRGB-D-SS perception module; (c) Wearable pRGB-D sensor; (d) Detection results for traversability awareness and water hazards avoidance.

Nowadays, unlike traditional approaches which detected water areas based on multi-feature fusion [9], Convolution Neural Networks (CNNs) learn and discriminate between different features directly from the input data by using a deeper abstraction of representation layers [5]. Notoriously, remarkable progress in most vision-based tasks have been fueled by the recently emerged deep learning pipelines and architectures. Semantic segmentation, as one of the challenging tasks that aims to partition an image into several

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 cofunded by State Key Laboratory of Modern Optical Instrumentation. This work has also been partially funded by the Spanish MINECO/FEDER through the SmartElderlyCar project (TRA2015-70501-C2-1-R) and from the RoboCity2030-III-CM project (Robótica aplicada a la mejora de la calidad de vida de los ciudadanos, fase III; S2013/MIT-2748), funded by Programas de actividades I+D (CAM) and cofunded by EU Structural Funds.

[1]Kailun Yang, Juan Wang and Kaiwei Wang are with College of Optical Science and Engineering, Zhejiang University, Hangzhou, China {elnino, zjuoptwj, wangkaiwei}@zju.edu.cn; [2]Luis M. Bergasa, Eduardo Romera and Elena López are with Department of Electronics, University of Alcalá, Madrid, Spain luism.bergasa@uah.es, eduardo.romera@edu.uah.es, elena@depeca.uah.es.

coherent semantically meaningful parts, is the key enabler to cover navigation-related perception needs in a unified way [12]. Intuitively, the detection of traversable area and water area could benefit from semantic segmentation because it directly leads to pixel-wise understanding and allows to exploit their inter-relations and contexts without imposing any assumptions. However, deep learning is data hungry and has not been integrated well with prior knowledge [13]. As far as water hazard is concerned, existing large-scale scene parsing datasets [14][15][16] only contain the classes of large water areas such as sea, river, lake and pool while annotated water puddles are lacking.

Apart from color and depth, polarization and its imaging extend information dimension to be used for material discrimination and target detection [17]. Given that light with different polarization states behave differently at the interface of objects surface, surface characteristics are coded implicitly as for materials, geometry structures and roughness. In this point of view, polarization attributes provide description of complementary surface features that can not be offered by color images. Light reflected from water surfaces is also polarized [10], so the utilization of polarization information for detecting it has obvious appeal, as perception systems should match natural human capacity to reach higher level of assistance for pedestrians with visual disability.

Based on above analysis, we propose the pRGB-D-SS perception framework (see Fig. 1), where Polarization imaging (p), RGB-D sensory awareness and real-time Semantic Segmentation (SS) are incorporated, creating pipelines to detect traversable areas and water hazards simultaneously. This paper considerably extends what was presented in [18], where we made the first attempt to detect water puddles for visually impaired pedestrians. Beyond the traversability awareness, we include novel contributions that reside in the following main aspects:

- A wearable pRGB-D sensor for navigational assistance in visually impaired individuals.
- A pRGB-D-SS perception framework which unifies the detection of water hazards including large water areas and small water puddles.
- A real-time semantic segmentation architecture to learn both local textures and global scene contexts without imposing any assumptions.

The remainder of this paper is structured as follows. In Section II, the framework is fully described in terms of the pRGB-D sensory awareness and the SS architecture. In Section III, the approach is evaluated and discussed as for real-time and real-world performance. Section IV draws the conclusions and gives an outlook to future work.

II. APPROACH

A. Polarized RGB-D sensory awareness

The overview of the pRGB-D-SS perception is depicted in Fig. 1, where the wearable pRGB-D sensor comprises a stereo camera, which is retrofitted by attaching horizontal and vertical polarization filters on the left and right camera

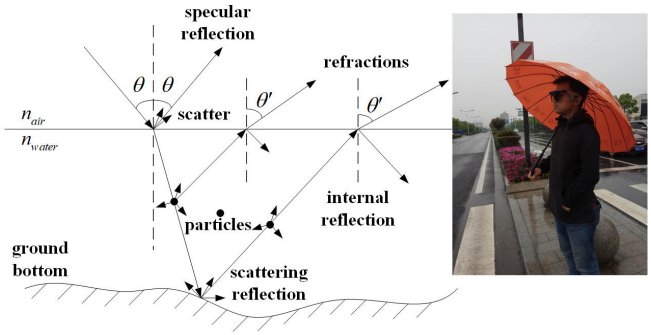


Fig. 2. The light reflection and refraction for water hazards with particles and ground bottom.

respectively. The stereo camera captures real-time RGB-D streams and transfer them to the processor, while the RGB images are fed into the segmentation network to obtain pixel-wise semantic information. A guided filter introduced in previous work [2] is utilized to enhance the original depth image to deliver large-scale depth information. After that, we warp the right image to the left image to produce point correspondences by using the disparity information that can be directly generated from the dense depth image. To construct a representation of polarimetric information, the polarized stereo pair is exploited to calculate a pixel-wise brightness difference image that indicates the degree of polarization resulting from reflection. For our purpose, this biologically inspired polarization-difference technique could be leveraged to aid navigation in visually impaired individuals, allowing us to form a pRGB-D-SS perception module as shown in Fig. 1(b), which supposes a very rich source of processed information, including semantic information, polarimetric information and enhanced depth information.

While this module could be utilized in many applications such as target detection [17], we focus on the study of hazard awareness. For water puddle detection, the primary cue is the polarization effect as specular reflection on water is known to polarize light [10]. The specular reflection from the water surface $R_{reflect}$ is the sum of two polarization components $R_{reflect,\perp}$ and $R_{reflect,\parallel}$, perpendicular and parallel respectively to the plane formed by the incident and reflected rays as given in (1) and (2). As marked in Fig. 2, n_{air} and n_{water} are the refractive indexes of air and water respectively and θ is the reflection angle at the water surface.

$$R_{reflect,\perp}(n_{air}, n_{water}, \theta) = \left[\frac{n_{air}\cos\theta - n_{water}\sqrt{1-(n_{air}/n_{water})^2\sin^2\theta}}{n_{air}\cos\theta + n_{water}\sqrt{1-(n_{air}/n_{water})^2\sin^2\theta}}\right]^2 \quad (1)$$

$$R_{reflect,\parallel}(n_{air}, n_{water}, \theta) = \left[\frac{-n_{water}\cos\theta + n_{air}\sqrt{1-(n_{air}/n_{water})^2\sin^2\theta}}{n_{water}\cos\theta + n_{air}\sqrt{1-(n_{air}/n_{water})^2\sin^2\theta}}\right]^2 \quad (2)$$

The polarized light from the air is supposed to comprise energy component $E_{\perp}^{S}(\theta)$ and $E_{\parallel}^{S}(\theta)$ for perpendicular and parallel components respectively as functions of reflection angle. The total energy entering the water can be calculated using (3). However, part of the energy F^S is scattered by suspended particles and ground bottom while the rest is

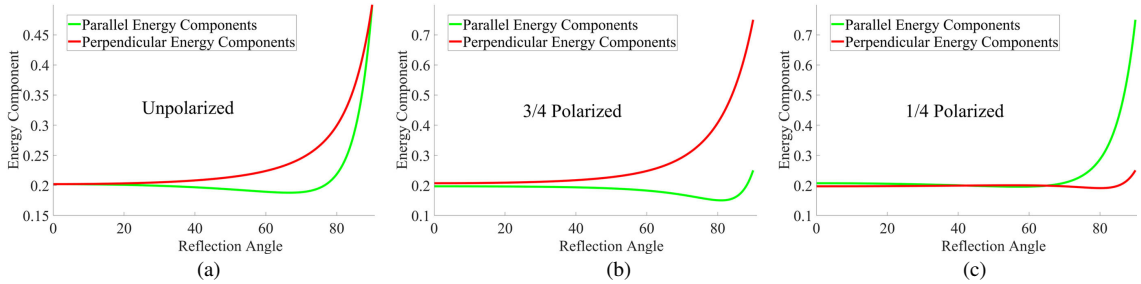


Fig. 3. Reflection and refraction energy components from water as function of degree and direction of polarization: (a) Unpolarized light; (b) Light is three quarters polarized in perpendicular direction with the water plane; (c) Light is a quarter polarized in perpendicular direction with the water plane.

absorbed by both particles and the ground as explained through (4) where $\mu_{particles}$ and μ_{bottom} are the scattering coefficients of particles and the ground bottom respectively, and $\mu_{absorption}$ is the absorption coefficient.

$$F^S = E^S_{\perp}(\theta)[1 - R_{reflect,\perp}(n_{air}, n_{water}, \theta)] + E^S_{\parallel}(\theta)[1 - R_{reflect,\parallel}(n_{air}, n_{water}, \theta)] \quad (3)$$

$$\mu_{particles} + \mu_{bottom} + \mu_{absorption} = 1 \quad (4)$$

Light in water can be considered as highly unpolarized light with random scattering and internal reflection. With part of the scattered light coming out of the water through refraction, the total light energy component coming out of the water is the summation of reflection and refraction for each polarization component, which can be calculated through (5) and (6) where θ' is the refraction angle from water to air.

$$E^R_{\perp}(\theta) = E^S_{\perp}(\theta)R_{reflect,\perp}(n_{air}, n_{water}, \theta) + 0.5F^S[\mu_{particles} + \mu_{bottom}]R_{refract,\perp}(n_{air}, n_{water}, \theta' = \theta) \quad (5)$$

$$E^R_{\parallel}(\theta) = E^S_{\parallel}(\theta)R_{reflect,\parallel}(n_{air}, n_{water}, \theta) + 0.5F^S[\mu_{particles} + \mu_{bottom}]R_{refract,\parallel}(n_{air}, n_{water}, \theta' = \theta) \quad (6)$$

For illustrative purposes, in Fig. 2, the water refractive index n_{water} is set to 1.33 as in most situations and the absorption coefficient $\mu_{absorption}$ is set to 60%. Fig. 3 shows that in both conditions, the polarization difference between perpendicular and parallel components is large enough to provide a strong cue for water hazards at reflection angles above 70 degrees or at distances above the minimum detection range. Based on this notion, it is able to detect water hazards by appropriate thresholding of the polarization difference with point correspondence from left color image to right color image. Following the rationale, water puddles are adaptively detected out of the traversable area by using the scheme presented in previous work [18]. In this regard, such fusion of polarimetric and semantic information is straightforward, which extends the CNN-based semantic segmentation by integrating the prior knowledge that hazardous puddles are encompassed by traversable areas.

B. Real-time semantic segmentation architecture

Up until very recently, pixel-wise semantic segmentation was not usable in terms of speed. However, a fraction of networks has focused on the efficiency by proposing architectures that could reach near real-time segmentation [19][20]. These advances have made possible the utilization of full scene segmentation in time-critical cases like blind assistance. To leverage the success of segmenting a variety of scenes and maintaining the efficiency, we design the architecture according to the SegNet-based encoder-decoder architectures like ENet [19] and our previous ERFNet [20]. In FCN-like architectures, feature maps from different layers need to be fused to generate a fine-grained output. As indicated in Fig. 4, our approach contrarily uses a more sequential architecture based on a encoder producing down-sampled feature maps and a subsequent decoder that up-samples the feature maps to match input resolution. In addition, Table I gives a detailed description of the integrated architecture. Currently, the residual layer adopted in state-of-art networks has two instances: the bottleneck version and the non-bottleneck design. In our previous work [20], "Non-bottleneck-1D" (non-bt-1D) was proposed, which is a redesign of the residual layer to strike a rational balance between the efficiency of the bottleneck and the learning capacity of non-bottleneck, by using 1D factorizations of the convolutional kernels. Thereby, it enables an efficient use of minimized amount of residual layers to extract feature maps and achieve semantic segmentation in real time.

However, for robust segmentation of traversable areas and water regions, we attach a different decoder with respect to the previous work. This critical modification aims to collect more contextual information while minimizing the sacrifices of learning textures. Global context information is of cardinal significance for navigational assistance in order to prevent delivering confusing feedback. To detail this, if the network mis-predicts a safe path in front of a large water area, the visually impaired would be left vulnerable in the dynamic environments. These risks could be mitigated by exploiting more context and learning more relationship between categories. With this target in mind, we reconstruct the decoder architecture. In this reconstruction, the decoder architecture follows the pyramid pooling module as introduced by PSPNet [21]. This module is applied to harvest different sub-region representations, followed by up-sampling and concatenation layers to form the final feature representations. For this reason, it carries both local and global context information from the pooled representations at different locations. Since it fuses features under a group of different pyramid levels, the output of different levels in this pyramid pooling module contains the feature map from the encoder with varied sizes. To maintain the weight of global feature, we utilize a convolution layer after each pyramid level to reduce the dimension of context representation to 1/*N* of the original one if the level size of the pyramid is *N*. As for the situation in Fig. 4c, the level size *N* equals

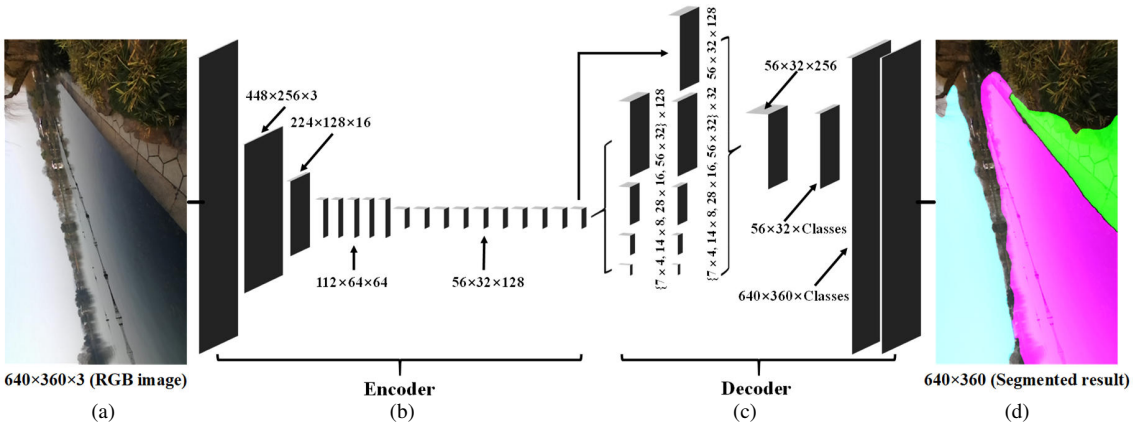


Fig. 4. The proposed architecture. From left to right: (a) Input, (b) Encoder, (c) Decoder, (d) Prediction.

to 4 and we decrease the number of feature maps from 128 to 32. Subsequently, the low-dimension feature maps are directly up-sampled to obtain the same size features as the original feature map through bilinear interpolation, such that the final per-pixel semantic predictions are fused with the polarization-based segmentation results at decision level to detect water puddles out of traversable paths.

TABLE I
LAYER DISPOSAL OF OUR PROPOSED NETWORK.
"OUT-F": NUMBER OF FEATURE MAPS AT LAYER'S OUTPUT,
"OUT-RES": OUTPUT RESOLUTION FOR INPUT SIZE OF 640×360.

	Layer	Type	Out-F	Out-Res
ENCODER	0	Scaling 640×360	3	448×256
	1	Down-sampler block	16	224×128
	2	Down-sampler block	64	112×64
	3-7	5×Non-bt-1D	64	112×64
	8	Down-sampler block	128	56×32
	9	Non-bt-1D (dilated 2)	128	56×32
	10	Non-bt-1D (dilated 4)	128	56×32
	11	Non-bt-1D (dilated 8)	128	56×32
	12	Non-bt-1D (dilated 16)	128	56×32
	13	Non-bt-1D (dilated 2)	128	56×32
	14	Non-bt-1D (dilated 4)	128	56×32
	15	Non-bt-1D (dilated 8)	128	56×32
	16	Non-bt-1D (dilated 2)	128	56×32
DECODER	17a	Original feature map	128	56×32
	17b	Pooling and convolution	32	56×32
	17c	Pooling and convolution	32	28×16
	17d	Pooling and convolution	32	14×8
	17e	Pooling and convolution	32	7×4
	17	Up-sampler and concatenation	256	56×32
	18	Convolution	C	56×32
	19	Up-sampler	C	640×360

III. EXPERIMENTS AND DISCUSSION

Experiments setup. The experiments are performed in public spaces around Westlake, the Zijingang Campus, the Yuquan Campus, the City College at Zhejiang University in Hangzhou and Venice Beach in Los Angeles. We captured real-world scenes wearing two navigation assistance systems including the smart glasses commercially available at *http://krvision.cn* and the customized pRGB-D prototype with 3D-printed shell which holds the sensors. In this fashion, two large-scale egocentric vision datasets are collected that can be accessed at *http://wangkaiwei.org/projecteg.html* including the terrain awareness dataset and the pRGB-D dataset. The metrics reported in this paper correspond to Intersection-over-Union (IoU), Pixel-wise Accuracy (P-A), Frame-level Accuracy (F-A) and Expansion Error (E-Error) that are prevailing in semantic segmentation challenges [14][15] and navigation tasks [1][2][10][18].

Real-time performance. The total computation time of a single frame at the resolution of 640×360 is 31ms, while the image acquisition takes 3ms, and the time costs for pRGB-D sensory awareness and SS are respectively 12ms and 16ms. In this sense, the computation cost is saved to maintain a reasonably qualified refresh-rate of 32.3FPS on a processor with a single cost-effective GPU GTX 1050Ti and a Core i7-7700HQ CPU. This inference time demonstrates that it is able to run our approach in real time, while allowing additional time for acoustic feedback [1][2][18] in navigation assistance or closed-loop control [4] in mobile robotics. Our ERF-PSPNet inherits the encoder design but implements a quite efficient version of decoder. Accordingly, the speed is even slightly faster than our previous approach with ERFNet, which runs at 29.4FPS on the same processor. In addition, on a embedded GPU Tegra TX1 (Jetson TX1) that enables higher portability while consuming less than 10 Watts at full load, our approach achieves approximately 14.1FPS.

Training setup. The challenging ADE20K [14] is chosen as it contains traversability-related classes and different water areas. To enrich images of different scenarios, we add 8733 images from PASCAL-Context [15] and 8132 images from COCO-Stuff [16] to obtain 37075 images in total. Additionally, we have 2000 images for validation from ADE20K. The most frequent 22 classes of terrains or objects are exploited for training while the water, sea, river, pool and lake are merged into a class of water hazards. To robustify the model against the real world, a group of data augmentations are performed including horizontally flipping with a 50% chance, jointly use of random cropping and scaling to resize the cropped regions into 448×256 input images. Random rotation by sampling distributions from the ranges [-20^o, 20^o] and color jittering from the ranges [-0.2, 0.2] for hue, [0.0, 2.0] for sharpness, [0.8, 1.2] for brightness, saturation and contrast are also applied. Our model is trained using Adam optimization, initiated with a batch size of 12, and a learning rate of 5×10^{-5} that decreases exponentially across epochs. Following the weight determining scheme in [19]

and the pretraining setup in [20], the training of the full network reaches convergence when cross-entropy loss is used as the criterion.

Segmentation accuracy. The accuracy of semantic segmentation is firstly evaluated on the challenging ADE20K dataset [14] by comparing the proposed ERF-PSPNet with deep neural networks in the state of the art for real-time segmentation including ENet [19] and our previous ERFNet [20]. Table II(a) details the accuracy of traversability-related classes including floor, road, grass, sidewalk, ground, the water areas and the mean IoU value. Since the classification of sky and water areas has the chance to be misled by the network that mainly relies on the local textures, the IoU of sky is analyzed together with the navigation-related classes. It could be told that the accuracy of most classes obtained with the proposed ERF-PSPNet exceeds the existing architectures that are also designed for real-time applications. Our architecture builds upon previous work but has the ability to collect more contextual information without major sacrifice of learning from textures. Accordingly, only the accuracy of sky is slightly lower than ERFNet.

TABLE II
ACCURACY ANALYSIS.
"P-A": PIXEL-WISE ACCURACY, "F-A": FRAME-LEVEL ACCURACY.

Network	Sky	Floor	Road	Grass	Sidewalk	Ground	Water	Mean
ENet [19]	89.7%	72.4%	69.4%	56.5%	38.2%	75.0%	67.3%	66.9%
ERFNet [20]	**93.2%**	77.3%	71.1%	64.5%	46.1%	76.3%	67.9%	70.9%
ERF-PSPNet	93.0%	**79.6%**	**75.6%**	**70.1%**	**51.3%**	**79.0%**	**78.7%**	**75.3%**

(a) On ADE20K dataset.

Approach	IoU	P-A	0-2m	2-3m	3-5m	5-10m
3D-RANSAC-F [1]	50.1%	67.2%	53.9%	91.8%	85.2%	61.7%
ENet [19]	62.4%	85.2%	79.9%	84.3%	89.7%	93.1%
Our ERF-PSPNet	**82.1%**	**93.1%**	**96.0%**	**96.3%**	**96.2%**	**96.0%**

(b) On terrain awareness dataset in terms of traversability awareness.

Accuracy	Sky	Traversability	Ground	Sidewalk	Water
IoU	88.0%	82.1%	72.7%	55.5%	69.1%
P-A	95.3%	93.1%	81.2%	93.1%	86.3%
0-2m	N/A	96.0%	76.9%	95.0%	96.2%
2-3m	N/A	96.3%	81.7%	96.5%	82.3%
3-5m	N/A	96.2%	87.4%	94.5%	76.9%
5-10m	N/A	96.0%	86.6%	93.6%	84.3%

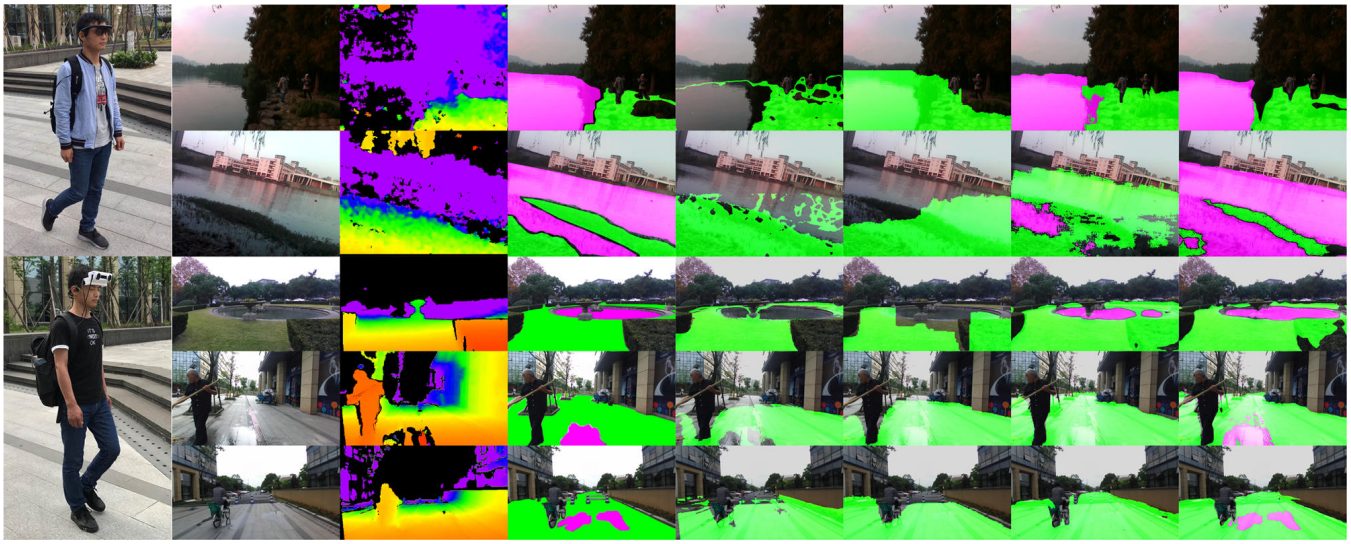
(c) ERF-PSPNet on terrain awareness dataset.

Approach	F-A of traversability	F-A of water	E-Error
3D-RANSAC-F [1]	79.6%	N/A	41.5%
3D-RANSAC-E [2]	93.8%	N/A	65.2%
3D-Tracking-P [10]	N/A	86.5%	N/A
FreeSpaceParse [3]	94.4%	N/A	60.6%
FreeSpaceParse-P [18]	94.4%	89.2%	**7.3%**
Our RGB-D-SS	**98.9%**	N/A	69.7%
Our pRGB-D-SS	**98.9%**	**92.4%**	8.1%

(d) On pRGB-D dataset in terms of traversability and water puddle detection.

Real-world performance. To analyze the major concern of detection performance for real-world assistance, we collect results over several depth ranges: within 2m, 2-3m, 3-5m and 5-10m on the terrain awareness dataset, which contains 120 images for testing with fine annotations of 7 important classes for navigation assistance including: sky, ground, sidewalk, stairs, water hazards, person and cars. This adequately considers that in navigational assistance, the short-range of ground area detection helps to determine the most walkable direction while superior path planning could be supported by longer traversability awareness [2]. Table II(b) shows both the IoU and the pixel-wise accuracy of traversability awareness, which is the cornerstone of navigational assistance. We compare the traversable area detection of our ERF-PSPNet to a state-of-the-art architecture ENet and a depth based segmentation approach 3D-RANSAC-F [1], which estimates the ground plane based on RANSAC and filtering techniques by using the dense disparity map. As the depth information of the ground area may be noisy and missing in high dynamic scenarios, we implemented a RGB image guided filter [2] to fill holes before random sampling of the 3D point cloud. In this way, the traditional 3D-RANSAC-F achieves decent accuracy ranging from 2m to 5m and it excels ENet from 2m to 3m as the depth map within this range is quite dense thanks to the active stereo design of the smart glasses. Still, our ERF-PSPNet outperforms ENet and 3D-RANSAC-F in both ranges.

For the visually impaired, it is preferred to know that there are water areas in some direction even if the segmented shape is not exactly accurate. This demand has been well satisfied as inferred from Table II(c) that the mean value of pixel-wise accuracy for traversable/water areas across different ranges is 90.5% and it achieves more than 96% within 2m in terms of water hazards. For the segmentation of water puddles vs. traversable areas, a distinction that is hard for humans to make consistently, we follow the metric in [1][2][10][18] to evaluate on a sequence of 1285 frames from the pRGB-D dataset. As illustrated in Table II(d), our pRGB-D-SS approach delivers high detection accuracy of traversability and puddles compared with respect to other works. Moreover, the expansion error is relatively low, illustrating the reliability of our approach, which seldom recognizes hazardous obstacles or water puddles as safe traversable area. It is worth mentioning that FreeSpaceParse [3], a procedure that renders Stixel-level segmentation with the original purpose for representing traffic situations, has been applied successfully thanks to the sensor fusion [18] by utilizing attitude angles and our polarization information. However, the procedure tailored to the problem relies on additional IMU observations and cannot handle large water areas. Our vision-based approach pursues the unified detection without requirements for attitude estimation, and achieves higher detection accuracy and competitive expansion error. Intriguingly, both the FreeSpaceParse and our pRGB-D-SS approach prove that polarimetric information are extremely important for water hazard avoidance and safety-critical traversability awareness, which cannot be guaranteed by RGB-D-based approaches. Fig. 5 exhibits the montage of pixel-wise results generated by our pRGB-D-SS approach, ENet, FreeSpaceParse and 3D-RANSAC-F. Qualitatively, our approach not only yields longer and more consistent segmentation which will definitely benefit the traversability awareness, but also retains the outstanding ability to unify the detection of large water areas and small water puddles within this framework.



(a) Prototype (b) RGB image (c) Depth image (d) Annotation (e) 3D-RANSAC-F (f) FreeSpaceParse (g) ENet (h) Our approach

Fig. 5. Qualitative examples of the segmentation on real-world images produced by our approach compared with ground-truth annotation, 3D-RANSAC-F [1], FreeSpaceParse [3] and ENet [19]. From left to right: (a) Wearable navigation systems including the commercial smart glasses and our customized prototype, (b) RGB image, (c) Depth image, (d) Annotation, (e) 3D-RANSAC-F, (f) FreeSpaceParse, (g) ENet, (h) Our pRGB-D-SS approach.

IV. CONCLUSION AND FUTURE WORK

In this contribution, we introduce a pRGB-D-SS perception module that incorporates polarized imaging, RGB-D sensor and real-time semantic segmentation. Based on this novel concept of perception, we unify the detection of water hazards including large water areas and small water puddles, and promote the awareness of traversability to aid navigation in visually impaired individuals. The proposed approach has been integrated in two wearable navigation systems, as well as evaluated on a large-scale challenging dataset and two egocentric real-world datasets, demonstrating the effectivity and applicability for navigation assistance.

Future works will involve the prediction of depth and polarization information from monocular images, as well as closed-loop field tests with real visually impaired users.

REFERENCES

[1] A. Rodríguez, J. J. Yebes, P. F. Alcantarilla, L. M. Bergasa, J. Almazán and A. Cela, "Assisting the visually impaired: obstacle detection and warning system by acoustic feedback," *Sensors*, 2012, 12(12), pp. 17476-17496.

[2] K. Yang, K. Wang, W. Hu and J. Bai, "Expanding the Detection of Traversable Area with RealSense for the Visually Impaired," *Sensors*, 2016, 16(11), 1954.

[3] H. C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré and D. Rus, "Enabling Independent Navigation for Visually Impaired People through a Wearable Vision-Based Feedback System," *IEEE International Conference on Robotics and Automation*. IEEE, 2017, pp. 6533-6540.

[4] T. Suleymanov, L. M. Paz, P. Piniés, G. Hester and P. Newman, "The path less taken: A fast variational approach for scene segmentation used for closed loop control," In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 3620-3626.

[5] R. A. Zeineldin, K. Saleh, M. Hossny, S. Nahavandi and N. A. El-Fishawy, "Navigational Path Detection for the Visually Impaired using Fully Convolutional Networks," *IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2017.

[6] H. Harms, E. Rehder, T. Schwarze and M. Lauer, "Detection of ascending stairs using stereo vision," In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 2496-2502.

[7] R. Cheng, K. Wang, K. Yang, N. Long, W. Hu, H. Chen, J. Bai and D. Liu, "Crosswalk navigation for people visual impairments on a wearable device," *Journal of Electronic Imaging*, 2017, 26(5), 053025.

[8] B. Xie, H. Pan, Z. Xiang and J. Liu, "Polarization-based water hazards detection for autonomous off-road navigation," In *Mechatronics and Automation (ICMA), 2007 International Conference on*. IEEE, 2007, pp. 1666-1670.

[9] A. Rankin and L. Matthies, "Daytime water detection based on color variation," In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on. IEEE, 2010*, pp. 215-221.

[10] C. V. Nguyen, M. Milford and R. Mahony, "3D tracking of water hazards with polarized stereo cameras," *arXiv preprint arXiv:1701.04175*, 2017.

[11] M. Martinez, A. Roitberg, D. Koester, R. Stiefelhagen and B. Schauerte, "Using Technology Developed for Autonomous Cars to Help Navigate Blind People," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1424-1432.

[12] E. Romera, L. M. Bergasa and R. Arroyo, "Can we unify monocular detectors for autonomous driving by using the pixel-wise semantic segmentation of CNNs?" *arXiv preprint arXiv:1607.00971*, 2016.

[13] G. Marcus, "Deep learning: A Critical Appraisal," *arXiv preprint arXiv:1801.00631*, 2018.

[14] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *arXiv preprint arXiv:1608.05442*, 2016.

[15] R. Mottaghi, X. Chen, X. Liu, N. G. Cho, S. W. Lee, S. Fidler, et al., "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 891-898.

[16] H. Caesar, J. Uijlings and V. Ferrari, "COCO-Stuff: Thing and Stuff Classes in Context," *arXiv preprint arXiv:1612.03716*, 2016.

[17] X. Huang, J. Bai, K. Wang, Q. Liu, Y. Luo, K. Yang and X. Zhang, "Target enhanced 3D reconstruction based on polarization-coded structured light," *Optics Express*, 2017, 25(2), pp. 1173-1184.

[18] K. Yang, K. Wang, R. Cheng, W. Hu, X. Huang and J. Bai, "Detecting traversable area and water hazards for the visually impaired with a pRGB-D sensor," *Sensors*, 2017, 17(8), 1890.

[19] A. Paszke, A. Chaurasia, S. Kim and E. Culurciello, "Enet: A deep neural network architecture for real-time segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[20] E. Romera, J. M. Alvarez, L. M. Bergasa and R. Arroyo, "ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2018, 19(1), pp. 263-272.

[21] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, Pyramid scene parsing network, In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881-2890.