Computer Vision for the Visually Impaired: the Sound of Vision System

Simona Caraiman[1], Anca Morar[2], Mateusz Owczarek[3], Adrian Burlacu[1], Dariusz Rzeszotarski[3], Nicolae Botezatu[1], Paul Herghelegiu[1], Florica Moldoveanu[2], Pawel Strumillo[3], and Alin Moldoveanu[2]

[1]"Gheorghe Asachi" Technical University of Iasi, Romania, www.tuiasi.ro
[2]University Politehnica of Bucharest, Romania, www.upb.ro
[3]Lodz University of Technology, Poland, www.p.lodz.pl

Abstract

This paper presents a computer vision based sensory substitution device for the visually impaired. Its main objective is to provide the users with a 3D representation of the environment around them, conveyed by means of the hearing and tactile senses. One of the biggest challenges for this system is to ensure pervasiveness, i.e., to be usable in any indoor or outdoor environments and in any illumination conditions. This work reveals both the hardware (3D acquisition system) and software (3D processing pipeline) used for developing this sensory substitution device and provides insight on its exploitation in various scenarios. Preliminary experiments with blind users revealed good usability results and provided valuable feedback for system improvement.

1. Introduction

The development of aids for helping the visually impaired to perceive the environment, to orientate and navigate has been the subject of many research works in the past two decades [20]. The reported efforts to support the rehabilitation of visually impaired have been directed towards the development of electronic travel aids (ETAs) and sensory substitution devices (SSDs). An ETA is a form of assistive technology with the purpose of enhancing mobility for the blind user [13, 20]. Sensory substitution devices are designed to convey visual information to the visually impaired by substituting visual information into one of their intact senses [1, 3, 5, 6, 15, 30, 35, 42]. These devices employ non-invasive human-machine interfaces, which, in the case of the blind, transform visual information into auditory or tactile representations using a predetermined transformation algorithm.

Although environment sensing techniques like sonar or radar have shown promising results, computer vision methods have more potential for providing an appropriate representation of the environment in real-world settings, which are noisy and difficult to interpret. Creating such a representation implies acquiring information and filtering it in order to provide the user with information that is not confusing and does not incur a sensory overload [22, 32, 33, 40, 48]. Moreover, as a general trend, higher quality image sensing devices are becoming cheaper, smaller and more widely available.

Analyzing the state of development for these assistive systems from the perspective of the end-user, we find that a plethora of works have been reported in the literature [4, 6, 9, 10, 12, 14, 16, 23, 27, 28, 34, 39, 41–45, 47, 49]. However, there are still some important steps to be taken before large communities of visually impaired users embrace this technology. The reasons for not having such consumer grade systems largely available for the end-user are related to many factors, such as form factor (wearability), lack of efficient training programs or general limitations of the visual rehabilitation itself.

Many computer vision-based assistive systems for the blind have tackled the problem of environment sensing and understanding [29], e.g. in the system reported in [31] semantic maps for indoor spaces were used to support high-level localization, navigation and context awareness. However, very few of the assistive systems consider the pervasiveness aspect [10, 27, 41] and work either in indoor or outdoor environments. This limitation mainly comes from the integrated 3D sensors. The infrared-based sensors, e.g., Kinect, do not cope with bright illumination from the Sun. Stereo sensors provide unreliable depth estimations in the presence of poor artificial lighting or uniformly colored/textured surfaces specific to indoor environments.

The system described by Kurata et al. [27] obtains positioning data from several sensing sources such as GPS, Wi-Fi, PDR (Pedestrian Dead Reckoning), image-based registration, and active RFID (if the infrastructure is in place),

and integrates them based on each uncertainty. Road-network data is also employed for map matching. Obstacle detection is performed using a laser range finder. An obstacle-map is rendered on a tactile display that is also used for Braille output.

The VeDi system [10] provides another showcase for indoor and outdoor navigation by integrating vision-based with pedestrian localization systems. The authors report a custom designed system that demonstrates how partially sighted people could be aided by the technology in performing an ordinary activity, like going to a mall and moving inside it to find a specific product. Computer vision techniques for detection, recognition and pose estimation of specific objects or features in the scene are combined with a hardware-sensor pedometer. Navigation in the indoor environment is performed using a Visual Navigator that searches for specific visual beacons (signs or environmental features).

Sensor fusion has also been exploited in the Navig project [24]. It uses GPS, two IMUs (Inertial Measurement Unit), one for body heading and a second one for head orientation, an adapted GIS and a stereo vision module. The vision module serves two functions: object localization and user positioning. The system looks for geolocated landmarks tagged in the GIS. When detected, these visual landmarks are not rendered to the user but are used to refine the current GPS position. The position estimate computed by the vision component is fused with GPS data to improve positioning. It can also be used in situations where the GPS positioning is faulty or not available.

While all these systems employ a form of data fusion from different sources, they only address the navigation problem and do not focus on the sensory substitution approach, or do not cope with any kind of environment. The Sound of Vision (SoV) system tackles the pervasiveness requirement by integrating both an IR-based depth sensor and a stereo vision system, together with an IMU device for recovering the head orientation. The main goal is to provide depth information in any environment (indoor or outdoor), in any illumination conditions and without the need for environment annotation. In this paper we present a prototype implementation of the SoVconcept, focusing on the computer vision component.

2. The Sound of Vision project

2.1. Description of the project and objectives

The Sound of Vision system is a non-invasive, wearable sensory substitution device that assists visually impaired people by creating and conveying an auditory and tactile (haptic) representation of the surrounding environment. This representation is created based on computer vision techniques, updated and conveyed to the blind users continuously and in real time. The objective of the SoV system is to aid both the perception and the navigation of visually impaired users in unknown environments. Moreover, the proposed system would provide feedback to the user in both indoor and outdoor environments and irrespective of the illumination conditions.

Figure 1: Acquisition devices support attached to different headgear designs. (Left) Rigid acrylic structure resembling VR headset implementations. (Right) Lightweight headgear with elastic strap bands.

The SoV system aims to aid the perception and mobility of visually impaired people who rely on assistive devices, such as the white cane, a guiding person or a guide dog in their daily lives. Thus, the visually impaired individuals that can benefit from the SoV system are those in Categories 3, 4 and 5 of visual impairment (according to the definition adopted by the World Health Organization [1]).

The work flow of the system consists in four main steps: (1) acquisition of 3D information from the environment, (2) 3D reconstruction of the sensed environment and segmentation into objects of interest, (3) audio and haptic modeling of the processed 3D scene, (4) rendering of audio and haptic information to the user.

The acquisition of 3D information from the environment is performed using depth cameras placed onto a rigid structure, which can be easily attached to various headgear designs (Figure 1). The acquisition system also integrates an IMU device that allows recovering the orientation of the head and cameras.

The 3D processing step performs a 3D reconstruction of the environment and identifies the elements of interest, such as ground, walls, ceiling, generic obstacles, negative obstacles (e.g., holes in the ground), doors, stairs, signs and texts.

The detected objects are further encoded using custom audio and haptic models. The system provides both full scene encodings and tools (modifiers of the main encodings). To open the possibility that the SoV system will be used by as many user categories as possible, including those with hearing impairments, the system provides both uni- and multi-modal encodings for scene description and is highly customizable. Moreover, the modifiers convey much simpler information without encoding all objects in the scene (e.g., flashlight mode, frontal selector, direction

[1] http://apps.who.int/classifications/icd10/browse/2015/en

of best navigable space in front of the user). Switching between the available encoding models and adjusting their audio/haptic options and volumes is easy to perform in real-time by the user, through a remote control.

Rendering the audio information is performed by means of several types of headphones: regular on/over headphones, in-ear headphones, and custom design multi-speaker headphones. The main requirement for the audio rendering unit is to be either open or hear-through, such that the visually impaired user is still able to perceive the natural sounds in the environment. Haptic information is conveyed to the user by means of a custom made belt, placed on the user's abdomen.

The SoV software runs on a portable computer carried in a custom made backpack with cooling facilities.

2.2. Requirements for the computer vision component

The main challenging requirements for the SoV system are: (i) to provide users with real-time feedback regarding the structure of the environment, (ii) to work in both indoor and outdoor environments, (iii) irrespective of the illumination conditions, (iv) to provide an added value compared to using the white cane, and (v) to be wearable. These general requirements translate into technical requirements for the computer vision components of the SoV device. They specifically have an important impact on the design of the 3D acquisition system and of the 3D processing pipeline.

While complex 3D processing algorithms need to be employed for the identification of the elements of interest in the environment, the system should still provide the user with real time feedback. To address these conflicting requirements, most of the 3D processing tasks are performed on the GPU. Moreover, the system adapts the detection of elements of interest based on the usage scenario. The most significant elements required in navigation scenarios are related to avoidance of obstacles and dangerous situations. In navigation scenarios, the user experiences a rapid change of scene structure, and thus a frequent change of elements encoded by the system. The aim is to keep the number and type of elements signaled by the system low enough, such that the user is able to understand the scene while moving. Some elements (e.g., doors, texts, signs) are only detected in scene exploration scenarios for which the encoding module of the system performs scanning of the reconstructed scene. This allows the 3D reconstruction module to perform complex and more time consuming algorithms for their detection. Furthermore, the detection of some elements (i.e., texts, signs, best free space) is explicitly triggered by the user.

In order to work in both indoor and outdoor environments, and irrespective of the illumination conditions, the 3D acquisition system employs two different types of depth sensors. Moreover, different 3D processing approaches are employed to deal with the specific structure and composition of indoor and outdoor environments, respectively.

The interplay between the SoV system and other assistive devices, such as the white cane, is envisioned and formulated based on recommendations from users and Orientation and Mobility (O&M) instructors. Thus, we expect the system to be used together with the white cane. We also expect that after some amount of training, the users would feel confident enough to use the SoV system without the white cane. Under these assumptions, the proposed system provides the user with both redundant and complementary information to the white cane. The SoV system is highly customizable by the user to adapt its output to the requirements corresponding to various scenarios of usage: with or without the white cane, simple/complex environments, various walking speeds, crowded/uncrowded environments.

3. 3D processing pipeline

3.1. Acquisition system

The acquisition system has a modular design with off-the-shelf components and employs fusion techniques to pre-process raw input data. It ensures depth data acquisition in multiple usage scenarios (e.g. indoor/outdoor, different lighting conditions).

3.1.1 Hardware

The acquisition devices are placed onto a rigid structure, which can be easily attached to various headgear designs. All the devices are connected to the SoV central processing unit via a USB 3.0 hub. These devices are: (1) A stereo RGB camera (SC) with configurable baseline (LI-OV580 from Leopard Imaging), used for outdoor image capture; (2) A Depth-of-Field camera (SS) (Structure Sensor PS1080 from Occipital), used for indoor and low light image capture; (3) An Inertial Measurement Unit (IMU) (LPMS-CURS2 from Lp-Research), used for tracking the head/camera orientation.

3.1.2 Operating modes

The 3D Acquisition system has four operating modes: Stereo camera input, Structure sensor input, Stereo-Structure dual input and recording playback. The first three modes are designed for the real-time use of the SoV system, whereas the playback mode is implemented for offline use with evaluation purposes. From a functional software perspective, the 3D Acquisition system is based on four distinct modules – one input module for each acquisition device (i.e. stereo camera, structure sensor, IMU device) and one main module for data synchronization, aggregation and preprocessing.

Stereo mode: The Acquisition module captures Stereo frames, synchronizes them with the IMU data, rectifies the left and right images and then applies a stereo correspondence algorithm (Elas [17] or SGBM [21]) in order to compute the disparity map.

Structure mode: The Acquisition module captures Structure frames (depth frames) and synchronizes them with the IMU data.

Dual mode: The Acquisition module captures Stereo frames, synchronizes them with the IMU data, rectifies the left and right images and then optionally runs a mapping procedure between RGB and depth frames or disparity and depth frames.

3.1.3 Sensor fusion

For the dual acquisition mode (i.e. projection of the depth output of the Structure sensor onto the RGB data from the Stereo camera) both devices must be calibrated together. The calibration process is performed with a custom developed calibration tool. It estimates the geometric distortions introduced by the cameras as well as the extrinsic parameters describing the transformation between each pair of cameras (left + right, left + IR, right + IR), where by IR we denote the infra-red sensor of the Structure Sensor device.

A "loose" synchronization is performed on the data from the three devices, captured on separate application threads. To this end we take into account the switching jitter of the acquisition threads occurring due to the load of the system and the OSs task switching mechanics. Experiments show that on a Windows 7 system with modest resources the jitter with both negative/positive values does not exceed 25% of the capture period at 30 fps and cancels itself after 10-15 frames. All captured data is timestamped (based on one of the system's steady clocks) and then a matching process is run based on the acquisition mode used.

The fusion between the output of the two imaging devices is necessary for the myriad of processing algorithms employed by the SoV system. The module implements two types of remapping: depth onto RGB and depth onto disparity. The remapping is performed based on the intrinsic and extrinsic parameters of the left RGB and IR sensors.

The depth image and the left rectified image can be fused together in the following way: (1) Recalculate the disparity values from stereo into depth values; (2) For each element in the depth map check if it is valid (has nonzero depth value). For valid elements, substitute an element in the disparity map from the stereo camera with the respective element from the reprojected depth map from the Structure Sensor. Valid elements from the Structure Sensor depth map have precedence over the corresponding elements in the stereo camera depth map.

3.2. Main processing pipeline

The main steps in the 3D processing pipeline are illustrated in Figure 2. The 3D processing module exploits different combinations of sensor data (Table 1) to maximize the system usability in different situations and still provide environmental information in conditions atypical to standalone sensors.

3.2.1 Indoor environments

The main indoor processing pipeline obtains depth images from the Structure Sensor and information about the camera rotation from the IMU. A point cloud is obtained from the depth map and the camera's intrinsic parameters. Most of the man-made objects from indoor environments, such as tables and chairs, have planar surfaces. This led to the idea of segmenting the 3D acquired point cloud into planar regions, which have similar normal vectors throughout all their contained points. After the segmentation, the obtained surfaces are merged into objects, based on the information from the IMU and on inter-frame consistency.

The *3D Reconstruction* stage consists of point cloud construction, edge detection, normal estimation and filtering. The 3D points from the point cloud have a one to one correspondence with the pixels in the depth map. Therefore, the normal vector is estimated as cross product over tangential vectors from the vicinity of the current point in image space. A rejection filter is next applied on the normal candidates, based on the deviation from an average normal computed inside a certain neighborhood. The final normal vector is obtained as a Gaussian weighted sum of the remaining normals from the neighboring window [36].

During the *segmentation*, a region growing step connects the pixels with similar properties in a scan-line traversal. For the current pixel, the regions containing its upper-left, upper, upper-right and left neighbors are inspected. A similarity cost is computed for the current pixel, with each of these regions, based on the difference in depth and normal. Next, a region merging step connects surfaces with similar statistics. Two regions are connected if they have approximately the same orientation, and the same distance from the camera to the theoretical planes containing the surfaces. The reasons behind introducing these similarity measures are explained in more detail in [36].

The *ground detection* is a very important step in this pipeline, because most of obstacle detection heuristics depend on a good estimation of the ground equation. A ground surface is horizontal in world coordinates, therefore its normal is approximately $[0, 1, 0]$. Also, the distance between the camera and the hypothetical plane containing such a surface is larger than the distance between the camera and other horizontal surfaces. The algorithm also uses inter-frame consistency in order to relax conditions for surfaces

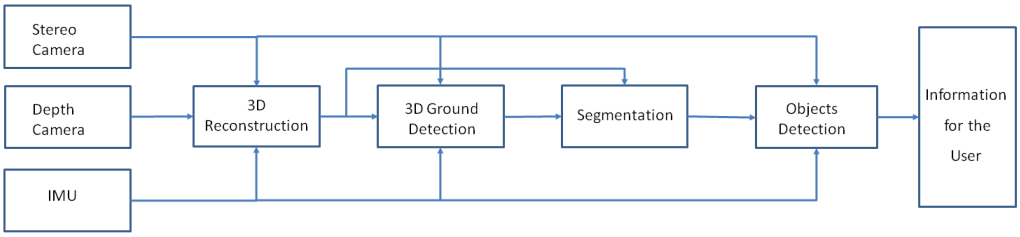

Figure 2: Main processing steps of the proposed system. The functionality and implementation of each step depends on the type of sensed environment (indoor vs. outdoor) and lighting conditions.

Table 1: The use of the data provided by the 3D acquisition module in different environments and illumination conditions

Environment	Lighting conditions	Input		Main approach
		Structure sensor	Stereo camera	
Indoor	Normal light	Depth Map	Left color map	Detect planar surfaces in the depth map and combine them into objects: detection of doors, texts,signs is performed on the color map
	Low light to complete darkness	Depth map	-	Detect planar surfaces and combine them into objects
Outdoor	Normal light	-	Left and right color maps	Detect objects based on disparity map histogram segmentation: estimate camera movement based on stereo pairs to perform object tracking
	Low light to complete darkness	Depth map	-	Detect objects based on depth map histogram segmentation; object tracking is not performed

that were already labeled as ground in previous frames [37]. Other heuristics are applied based on the previously detected ground region, in order to remove false positives from the ceiling surface candidates. The *walls* are large surfaces perpendicular to the ground. Similar to the heuristics for the ground, the wall detection algorithm employs inter-frame consistency in order to relax the conditions for surfaces that have been labeled as walls in previous frames. The remaining surfaces are merged into generic objects based on adjacency [37].

Since an indoor environment can contain many objects, the user might get disoriented if all these objects would be encoded. Therefore, in a configuration step, the user chooses how many objects should be encoded and how to choose the most relevant ones, based on the size, depth and deviation from the view direction. The size of an object S is determined as a product of its width and height. The deviation from the view direction of an object, D is computed in image space as the difference between the X coordinate of the image center and the average X coordinate of the object. For each object, an importance cost is computed as a weighted sum of costs given by the object's size, depth and deviation from the view direction: $C = C_{size} + C_{depth} + C_{dev}$. $C_{size} = S/S_{max}$, $C_{depth} = Z/Z_{min}$ and $C_{dev} = D/D_{min}$, where S_{max} is the size of the biggest object, Z and Z_{min} represent the depth of the current object, respectively the depth of the closest object to the camera and D_{min} is the deviation of the object located closest to the image center along the X direction. The objects with the biggest importance cost are sent to the encoding.

3.2.2 Outdoor environments

The outdoor processing pipeline exploits data from both the stereo camera and the IMU sensor for a 3D reconstruction approach where a global 3D model is built. The reconstruction of a global 3D model is necessary as it allows coping with the high amount of errors in the estimation of depth from disparity. The global 3D model is built using state of the art algorithms for disparity computation [17] and camera motion estimation [26]. The approach of independently segmenting each frame in the presence of these errors can lead to erroneous object detection and thus to unreliable functioning of the system. However, this approach is more reliable in low light conditions where the depth from the Structure Sensor is used as input.

A very important step in the detection of objects in the

environment is represented by the correct estimation of the ground surface. The ground surface is detected by first estimating the equation of a plane that approximates this surface [19]. Second, all the 3D points in the global model that fit this plane equation within a threshold are considered part of the ground surface. The threshold was empirically selected with a value of 15 cm, which accounts for both the usual slight unevenness of the real ground surfaces and for 3D point estimation errors. Moreover, uneven surfaces within this oscillation of level generally do not pose threats to the safety of VIP's mobility, especially when using the white cane.

The global 3D model is consistent along the time line of the system use. Consistency is achieved by incrementally adding the 3D representations of the individual frames in the stereo stream to this 3D model. A confidence measure is associated to each 3D point forming the global model. This confidence linearly depends on the number of frames in which the points could be tracked, i.e., the point was in the sensor's field of view and the disparity computation algorithm could provide a 3D measurement for it. While introducing the confidence measure for the 3D points greatly improves the accuracy of the estimation of static regions in the environment, it also intrinsically excludes the dynamic objects. To overcome this difficulty, we employ the use of fusion maps and color difference maps between the current frame and the previous global 3D model to determine 3D measurements corresponding to dynamic objects. To refine the dynamics estimation we correct both the false negative dynamic 3D points and the false positives ones using a statistical approach. We exploit a superpixel segmentation [2] of the color image and mark the superpixels as dynamic or static based on the votes of their corresponding 3D points.

In our approach, the 3D global model is segmented into objects after excluding the regions corresponding to the ground surface [7]. Objects are tracked between frames based on labels assigned to the 3D points in each frame. This ensures their temporal consistent identification throughout the stream. Moreover, the ground information is exploited for computing the best free space in front of the user and for the detection of negative obstacles on the pathway. The superpixels segmentation is used for refining the object segmentation as the superpixels are computed such as their boundaries generally do not cross object boundaries.

In low light outdoor scenarios, where the stereo sensor cannot be used, the system acquires depth information using the Structure Sensor. In these scenarios, the global 3D model approach is not used as we find that camera motion estimation algorithms that exploit only depth information do not provide the required reliability. The segmentation of outdoor depth images from the structure sensor is performed in a similar manner as in the case of stereo data, except for the exploitation of color information.

3.2.3 Detection of special objects in the environment

Negative obstacles can be represented by holes in the ground, potholes or any large difference of height between two ground surfaces (e.g., edge of a railway station platform, stairs down). The SoV system detects their presence based on empirical assumptions regarding their image characteristics. The reliability of this approach is improved by employing a tracking mechanism in order to validate the identified candidates [19].
Doors detection algorithm is working on color image, depth map and identified ground plane equation as input data. The method relies on detection of lines which are matched with a proposed geometric model of doors. Detecting and tracking the door handles allows us to reliably localize them, even if they are only partially visible in the scene [46].
Stairs. The indoor stairs detection algorithm clusters patches that have the normal vectors oriented vertically in world space. It also applies several filters to eliminate false positives, based on edges, surface rotation and distance to the camera, as suggested in [11].
Sign detection is carried out by means of a supervised classifier trained using sets of annotated images containing signs for exit, toilet, bus stop, pedestrian crossing. The classifier is an SVM (Support Vector Machine) which uses an fHOG-based approach for feature identification [25]. It was also trained to identify semaphores and differentiate between red and green light state.
Text detection is performed by integrating the *Class-Specific Extremal Regions* (CSER) [8, 38] method for text candidates localization together with a proven, reliable, and open-source *Tesseract OCR* library [18] developed by Google.

3.2.4 Computing the properties of detected objects and impact on user's perception

The list of objects and their properties is used by the audio and haptic encoders. With these encodings, the user receives information regarding the localization of objects, their size and type, as well as their elevation. A first round of usability experiments have been performed with 12 visually impaired and 12 sighted persons, using the system in custom generated virtual environments, of increasing complexity, and with a separation between the system's audio and haptic encodings. The conclusions of these experiments pointed out that, while the users could easily perceive different widths/heights of objects when presented with a static image, it is difficult for them to get accustomed with these measures variating when rotating the head. Such variations can be perceived when the width and height of objects are computed based on the projection of the objects onto the camera image plane (Figure 3). Moreover, in ego-dynamic testing scenarios, we found out that passing between two

obstacles can pose difficulties when the user is only informed about the position and width of the two objects. This task requires that the user calculates the navigable space between the obstacles based on their properties. We addressed these limitations by:

(1) Increasing the usability of object properties for navigability and scene perception of the visually impaired. Their perception is different from the perception of these properties by a sighted person. A sighted person interprets the width, height and length of a car the same way, irrespective of how he/she looks at it. Moreover, the perception of a sighted person depends on the orientation of the object with respect to his/her position and not to the head orientation. We adopt a definition that allows the user to interpret these properties with respect to the navigable space around him/her: the width of an object is considered to be the horizontal dimension of the space occupied by the object if the user was heading towards the center of the object, while the height is computed as the vertical one. The vertical and horizontal axes are defined with respect to the orientation of the ground. Figure 3 illustrates the computed width of an object using both methods, with two head orientations for the same user position.

(2) Conveying information regarding the best navigable space in front of the user. This information indicates the direction (azimuth) the user can navigate to and the depth of the free space in that direction (in meters). The radial slice of best free space is chosen to be the closest to the camera heading direction, in case multiple ones are detected. Moreover, a minimum width of free space is considered (0.9m), equivalent to a standard door opening. The slices of navigable space are computed based on radially sampling the detected ground surface and obstacles (hanging, on the ground and negative ones).

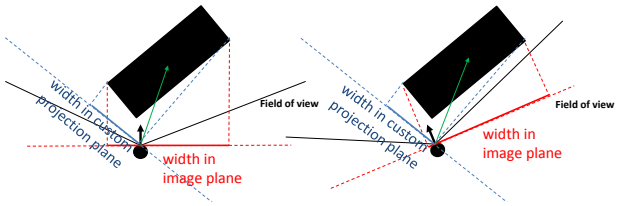

Figure 3: Computing the width of an object based on its projection on the camera plane (red) and custom plane (blue). The width computed with the proposed method (blue) is constant when variating the camera orientation from the same user position.

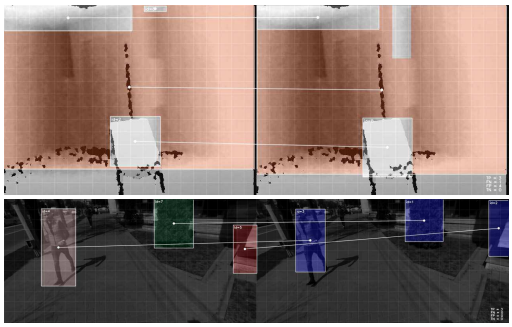

Figure 4: Example correspondence between objects within the ground truth (left) and the detection results (right) in the indoor (top) and outdoor (bottom) environments.

4. Evaluation

4.1. Technical performance

The evaluation of system accuracy is performed automatically, using a custom developed application. The main objective is to evaluate the output of the system from the user's perspective, i.e., with respect to the number, type, position and size of the encoded objects. To this end, we evaluate the segmentation performed by the system against up to 40 manually annotated ground truth (GT) images – RGB frames for outdoor and depth maps for indoor environment. For the evaluation of object size, we use the bounding boxes of the object projection onto the image plane, as the GT for the actual system output cannot be obtained. The evaluation is performed for 5 and 10 meters distance ranges, in indoor and outdoor scenarios. The correspondence (Figure 4) between objects in the GT image and those in the output image is established using a similarity metric based on the Jaccard similarity coefficient (JC). JC is defined as $\frac{area(R_{GT} \cap R_D)}{area(R_{GT} \cup R_D)}$, where R_D is the rectangular region of the detection result and R_{GT} denotes the ground truth. An object is considered detected properly (true positive) if the JC is larger than or equal to 50% (a score larger than 50% is normally considered as a "good"). The sensitivity ($TPR = TP/(TP + FN)$), positive predictive value ($PPV = TP/(TP + FP)$) and accuracy ($ACC = (TP + TN)/(TP + FP + FN + TN)$) of the system are presented in Table 2, where TP denotes true positive, FP – false positive, FN – false negative and TN – true negative detections, respectively. When checking whether the correspondence between objects is right, parameters such as width, height and distance between centers of rectangular regions are taken into account. The error in computing the width/height of objects is evaluated as a ratio of the difference between GT and computed values and of GT, only for the TP detections. The center deviation is also computed in image space using the average Euclidean distance between GT center and computed center. Ground surface detection is evaluated pixel wise.

Within the main problems of the outdoor pipeline are clustering of several objects into a single one and the fact that lower parts of objects are sometimes considered as

Table 2: Evaluation of the system performance

Environment Element	TPR	PPV	ACC
Ground Surface	0.98	0.90	0.89
Obstacles	0.97	0.78	0.76
	width error: 0.13m		
	height error: 0.17m		
	center deviation: 16px		

ground surface. Since our automatic evaluation procedure does not currently support splits and merges, those results may be wrongly marked as improper. Indoor pipeline, on the other hand, is characterized by over-segmentation of walls. In both pipelines we achieved very good results of ground surface detection despite a slight over-segmentation. It is very important as audio and haptic encoding of the scene (where ground is removed as safe) and object detection based on remaining object regions heavily rely on correct identification of ground surface.

The reconstruction and segmentation runs at approx. 15 fps for indoor environments and at approx. 10 fps for the outdoor pipeline, including the computation of negative obstacles and best free space. The times were measured on a consumer laptop (Intel Core i7-4720HQ Processor with a GTX 970M GPU). Stairs and door detection run at 10 fps, while signs detection runs at 20 fps. Text detection can be very time consuming, depending on scene complexity. Thus, it is not run in real time, but only when triggered by the user and associated with the scene scanning mode. If the module doesn't return a result within the system's scanning time (1.5 s), the system reports a failure in detecting any text in the environment.

4.2. Preliminary usability assessment

Preliminary usability experiments with the developed system have been carried out in two rounds of experiments. First experiments were conducted with 12 visually impaired participants (VIP) and 12 sighted persons. The tests were run with custom Virtual Environments (VE). The main objective was to assess the usability of several audio and haptic encodings. The results helped us improve both the encodings and the computation of object properties. The second round of tests were performed with 19 VIPs with the involvement of O&M instructors. The experiments were designed with increasing difficulty and were interspersed with training. There were two stages of the experiment, first in VE, followed by Real World (RW) setups. The tests involved both ego-static and ego-dynamic scenarios. The RW tests consisted in modeled indoor environments with cardboard boxes in random locations playing the role of obstacles. Besides the collected psychophysics measurements (e.g., accuracy, response times), we received rich feedback from the participants for further improvement of the prototype, both by means of validated questionnaires and personal interviews. An extensive report on the results of these usability experiments is not in the scope of the paper. However, there are a few conclusions worth mentioning with respect to the 3D acquisition and processing components of the system: (i) a more pleasant and comfortable headgear design should be devised for the final system implementation, (ii) RW tests were in general more difficult than the VE tests, however, provided satisfactory results as participants were in general able to complete the task objectives, (iii) while a 15fps update to the user is not in the standard understanding of real time operation in the computer animation field, it did not pose any problems to the real time perception of the environment in navigation scenarios, (iv) users are able to cope with regions of walls reported by the system as generic obstacles, even in scene exploration mode.

5. Conclusions

The Sound of Vision system is a complex sensory substitution device that heavily relies on computer vision techniques to convey environment information to visually impaired users. The objective of this paper was to provide an overview of the entire computer vision based system, emphasizing on how the conflicting user requirements (real time feedback, complex environments and lighting conditions) are addressed. The preliminary evaluation of the system was performed with respect to the accuracy of the produced output in normal functioning mode, considering the main categories of elements (ground, walls and generic objects). Additional components integrated in the 3D processing pipeline have been reported and evaluated individually. Further work will address more in depth evaluation and improvement, specifically with respect to system usability given its accuracy. That is, we are interested in evaluating how the system's FP and FN rates affect the user perception of the scene with respect to navigability and scene understanding. Preliminary experiments with visually impaired users have shown that, with a relatively small amount of training, they are able to use the system for scene understanding and obstacle avoidance. Further experimentation will be performed aiming at more detailed user feedback in more complex environments. For example, such experiments would confirm/infirm our supposition that under-segmentation would be preferred for areas with clustered objects, especially for mobility scenarios.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643636 "Sound of Vision".

References

[1] S. Abbouda, S. Hanassya, S. Levy-Tzedek, S. Maidenbaum, and A. Amedi. Eyemusic: Introducing a visual colorful experience for the blind using auditory sensory substitution. *Restorative Neurology and Neuroscience*, 32(2), 2014.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, Nov. 2012.

[3] M. Auvray, S. Hanneton, and J. K. O'Regan. Learning to perceive with a visuo - auditory substitution system: Localisation and object recognition with 'the voice'. *Systems Journal, IEEE*, 36(3):416–430, 2007.

[4] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, and S. Yaacob. A stereo image processing system for visually impaired. *International Journal of Signal Processing*, 2(3):136, 2008.

[5] G. Bologna, B. Deville, T. Pun, and M. Vinckenbosch. Transforming 3d coloured pixels into musical instrument notes for vision substitution applications. *EURASIP International Journal of Image and Video Processing*, 2007(2):1–15, 2007.

[6] M. Bujacz, P. Skulimowski, and P. Strumillo. Naviton - a prototype mobility aid for auditory presentation of three-dimensional scenes to the visually impaired. *Journal of the Audio Engineering Society*, 60(9):696 – 708, 2012.

[7] A. Burlacu, S. Bostaca, I. Hector, P. Herghelegiu, G. Ivanica, A. Moldoveanul, and S. Caraiman. Obstacle detection in stereo sequences using multiple representations of the disparity map. In *2016 20th International Conference on System Theory, Control and Computing (ICSTCC)*, pages 854–859, Oct 2016.

[8] D. Chen, J.-M. Odobez, and H. Bourlard. Text detection and recognition in images and video frames. *Pattern Recognition*, 37(3):595 – 608, 2004.

[9] L. Chen, B.-L. Guo, and W. Sun. Obstacle detection system for visually impaired people based on stereo vision. In *Genetic and Evolutionary Computing (ICGEC), 2010 Fourth International Conference on*, pages 723–726, Dec 2010.

[10] P. Chippendale, V. Tomaselli, V. D'Alto, G. Urlini, and C. Modena. Personal shopping assistance and navigator system for visually impaired people. In *Proc. of the CVPR2014 Workshop"*, 2014.

[11] A. Ciobanu, A. Morar, F. Moldoveanu, L. Petrescu, O. Ferche, and A. Moldoveanu. Real-time indoor staircase detection on mobile devices. In *International Conference on Control Systems and Computer Science (CSCS21)*, 2017.

[12] P. Costa, H. Fernandez, P. Martins, J. Barroso, and L. Hadjileontiadis. Obstacle detection using stereo imaging to assist the navigation of visually impaired people. *Procedia Computer Science*, 12:83 – 93, 2012.

[13] D. Dakopoulos and N. G. Bourbakis. Wearable obstacle avoidance electronic travel aids for blind: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1):25–35, 2010.

[14] L. Dunai, B. Garcia, I. Lengua, and G. Peris-Fajarnes. 3d cmos sensor based acoustic object detection and navigation system for blind people. In *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, pages 4208–4215, Oct 2012.

[15] L. Dunai, G. Peri-Fajarnes, E. Lluna, and B. Defez. Sensory navigation device for blind people. *The Journal of Navigation*, 66:349 362, 2013.

[16] V. Filipe and et. al. Blind navigation support system based on microsoft kinect. *Procedia Computer Science*, 14(0):94 – 101, 2012.

[17] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision (ACCV)*, 2010.

[18] I. Google. Tesseract open source ocr engine (main repository). [online]. Accessed on July 1st, 2017.

[19] P. Herghelegiu, A. Burlacu, and S. Caraiman. Robust ground plane detection and tracking in stereo sequences using camera orientation. In *2016 20th International Conference on System Theory, Control and Computing (ICSTCC)*, pages 514–519, Oct 2016.

[20] M. Hersh and M. Johnson. *Assistive Technology for Visually Impaired and Blind People*. Springer Publishing Company, 2008.

[21] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb 2008.

[22] R. Jafri, S. Ali, H. Arabnia, and S. Fatima. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. *Visual Computing*, 30:1197 – 1222, 2014.

[23] J. Jose, M. Farrajota, J. M. Rodrigues, and J. H. du Buf. The smartvision local navigation aid for blind and visually impaired persons. *JDCTA: International Journal of Digital Content Technology and its Applications*, 5(5):362 – 375, 2011.

[24] S. Kammoun, G. Parseihian, O. Gutierrez, A. Brilhault, A. Serpa, M. Raynal, B. Oriola, M.-M. Mac, M. Auvray, M. Denis, S. Thorpe, P. Truillet, B. Katz, and C. Jouffrais. Navigation and space perception assistance for the visually impaired: The {NAVIG} project. {*IRBM*}, 33(2):182 – 189, 2012. Numro spcial {ANR} {TECSANTechnologie} pour la sant et l'autonomie.

[25] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[26] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In *Intelligent Vehicles Symposium (IV)*, 2010.

[27] T. Kurata, M. Kourogi, T. Ishikawa, Y. Kameda, K. Aoki, and J. Ishikawa. Indoor-outdoor navigation system for visually-impaired pedestrians: Preliminary evaluation of position measurement and obstacle display. In *Wearable Computers (ISWC), 2011 15th Annual International Symposium on*, pages 123–124, June 2011.

[28] Y. H. Lee, T.-S. Leung, and G. Medioni. Real-time staircase detection from a wearable stereo system. In *21st International Conference on Pattern Recognition (ICPR 2012)*, pages 3770–3773, 2012.

[29] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. Farinella. Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154:1 – 15, 2017.

[30] S. Levy-Tzedek, D. Rimer, and A. Amedi. Color improves 'visual' acuity via sound. *Frontiers in Neuroscience*, 8(358), 2014.

[31] B. Li, P. Muoz, X. Rong, J. Xiao, and Y. T. nad Aries Arditi. Isana: Wearable context-aware indoor assistive navigation with obstacle avoidance for the blind. In *Lecture Notes in Computer Science*, volume 9914, pages 448–462, 2016.

[32] S. Maidenbaum, S. Abboud, and A. Amedi. Image and video processing for visually handicapped people. *Neuroschience and Biobehavioral Reviews*, 41:3 – 15, 2014.

[33] R. Manduchi and J. Coughlan. (computer) vision without sight. *Commun. ACM*, 55(1):96–104, Jan. 2012.

[34] S. Mattoccia and P. Macri. 3d glasses as mobility aid for visually impaired people. In *Proc. of the ECCV2014 Workshop*, 2014.

[35] F. L. M. Milottta, D. Allegra, F. Stanco, and G. M. Farinella. An electronic travel aid to assist blind and visually impaired people to avoid obstacles. In *International Conference on Computer Analysis of Images and Patterns*, pages 604–615, 2015.

[36] A. Morar, F. Moldoveanu, L. Petrescu, O. Balan, and A. Moldoveanu. Time-consistent segmentation of indoor depth video frames. In *International Conference on Telecommunications and Signal Processing*, 2017.

[37] A. Morar, F. Moldoveanu, L. Petrescu, and A. Moldoveanu. Real time indoor 3d pipeline for an advanced sensory substitution device. In *International Conference on Image Analysis and Processing*, 2017.

[38] L. Neumann and J. Matas. Text localization in real-world images using efficiently pruned exhaustive search. In *2011 International Conference on Document Analysis and Recognition*, pages 687–691, Sept 2011.

[39] E. Peng, P. Peursum, L. Li, and S. Venkatesh. A smartphone-based obstacle sensor for the visually impaired. In Z. Yu, R. Liscano, G. Chen, D. Zhang, and X. Zhou, editors, *Ubiquitous Intelligence and Computing*, volume 6406 of *Lecture Notes in Computer Science*, pages 590–604. Springer Berlin Heidelberg, 2010.

[40] T. Pun, P. Roth, G. Bologna, K. Moustakas, and D. Tzovoras. Image and video processing for visually handicapped people. *EURASIP Journal on Image and Video Processing*, pages 1 – 12, 2007.

[41] L. Ran, S. Helal, and S. Moore. Drishti: An integrated indoor/outdoor blind navigation system and service. In *IEEE International Conference on Pervasive Computing and Communications*, pages 23–32, 2004.

[42] F. Ribeiro, D. Florencio, P. Chou, and Z. Zhang. Auditory augmented reality: Object sonification for the visually impaired. In *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on*, pages 319–324, Sept 2012.

[43] A. Rodriguez, J. Yebes, P. Alcantarilla, L. Bergasa, J. Almazan, and A. Cela. Obstacle avoidance system for assisting visually impaired people. In *Proc. of the 2012 IEEE Intelligent Vehicles Symposium Workshops*, 2014.

[44] J. Saez, F. Escolano, and M. Lozano. Aerial obstacle detection with 3d mobile devices. *IEEE J Biomed Health Inform*, 19:74 – 80, 2015.

[45] J. M. Saez Martinez and F. Escolano Ruiz. Stereo-based Aerial Obstacle Detection for the Visually Impaired. In *Workshop on Computer Vision Applications for the Visually Impaired*, Marseille, France, Oct. 2008. James Coughlan and Roberto Manduchi.

[46] P. Skulimowski, M. Owczarek, and P. Strumillo. Door detection in images of 3d scenes in an electronic travel aid for the blind - in review. In *10th International Symposium on Image and Signal Processing and Analysis*, 2017.

[47] R. Tapu, B. Mocanu, A. Bursuc, and T. Zaharia. A smartphone-based obstacle detection and classification system for assisting visually impaired people. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 444–451, 2013.

[48] J. Terven, J. Salas, and B. Raducanu. New opportunities for computer vision-based assistive technology systems for the visually impaired. *Computer*, 47:52 – 58, 2014.

[49] Z. Wang, H. Liu, X. Wang, and Y. Qian. Segment and label indoor scene based on rgb-d for the visually impaired. In C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. OConnor, editors, *MultiMedia Modeling*, volume 8325 of *Lecture Notes in Computer Science*, pages 449–460. Springer International Publishing, 2014.