A Mobility Device for the Blind with Improved Vertical Resolution Using Dynamic Vision Sensors

Lukas Everding, Lennart Walger, Viviane S. Ghaderi, and Jörg Conradt
Neuroscientific System Theory, Technical University of Munich, 80333 München, Germany

Abstract—We propose an improved version of a wearable lightweight device to support visually impaired people during their everyday lives by facilitating autonomous navigation and obstacle avoidance. The system deploys two retina-inspired Dynamic Vision Sensors for visual information gathering. These sensors are characterized by very low power consumption, low latency and drastically reduced data rate in comparison with regular CMOS/CCD cameras which makes them well suited for real-time mobile applications.

Event-based algorithms operating on the visual data stream extract depth information in real-time which is translated into the acoustic domain. Spatial auditory signals are simulated at the computed origin of visual events in the real world. These sounds are modulated according to the position in the field of view which the user can change by moving their head. Here, different tests with eleven subjects are conducted to evaluate the performance of the system. These tests show that the modulation helps to improve object localization performance significantly in comparison to prior experiments. Further trials estimate the visual acuity a user of the device would have using the Landolt C test.

The low power consumption of all integrated components in a final system will allow for a long lasting battery life of a small portable device, which might ultimately combine perceived visual information and environmental knowledge to provide a higher quality of life for the visually impaired.

Index Terms—wearable assistive device, mobility aid, event-based vision, acoustic scene representation, sensory substitution

I. INTRODUCTION

Visually impaired people are faced with numerous challenges on a daily basis which can severely impact the quality of their private as well as their professional lives. A major problem is autonomous and independent orientation and movement. Although modern technology has found its way into most aspects of life, for the blind the traditional white cane is still the most common mobility aid. The information it delivers to the user is, however, very limited. Spatial and temporal resolution are coarse, it can only detect obstacles in very close proximity and no objects above hip level (like road signs or tree branches).

Recently developed Electronic Travel Aids (ETAs) aim to overcome the limitations of the white cane by creating a richer representation of a person’s surroundings and enabling them to move more freely and autonomously. They are non invasive wearable devices. Typically, ETAs are equipped with sensors (mostly cameras or ultrasound sensors) to collect spatial information about the environment and convey this information to the user via touch or hearing.

Different systems are available such as the vOICe [1], OIWOBB [2], BrainPort [3] or the GuideCane [4], which use different sensors or user interfaces:

- the vOICe uses an encoding scheme to convert the pixels of a grey scale picture where vertical position corresponds to sound frequency, horizontal position time delay and brightness to amplitude. It requires substantial training to become able to interpret the observed scene, but it was found that experienced users could retrieve visual information like letter orientation using this device [5].

- OIWOBB is a system that resembles the one suggested here most closely. It uses two regular cameras to capture visual information as well as an inertial measurement unit (IMU) to track head position and conveys position of obstacles and information about static background features using different sound patterns. Extracting depth and shape information from frame-based camera data requires elaborate calculations and makes miniaturization of the system difficult.

- Brainport uses the sense of touch to convey information to the user. Spatial information is collected using a camera, translated into pulse pattern and sent to the user’s tongue via electrodes. After some hours of training users were able to perform simple visual tasks like identifying line orientation or objects.

- The GuideCane uses active sensing with ultrasound to scan the environment and detect obstacles. The sensor is attached to a handle and equipped with wheels. In order to operate the device the user has to push it. As soon as an obstacle is detected the wheels change direction automatically to steer the user in a safe direction.

Although, these ETAs are available, they are not commonly used by visually impaired people. Possible reasons are low spatial or temporal resolution, long training times, non-intuitive and uncomfortable design, interference with user’s activities. We suggest a new portable system which translates visual information into virtual three dimensional sound which represents the environment as audio landscape [6]. This work presents a refined version of this device and addresses the limitations it previously had. An overview over the whole system can be seen in Fig. 1.

II. DEVICE

A. Concept of event-based vision

In our approach, Dynamic Vision Sensors (DVS, [7]) are used for gathering visual information instead of frame-based
cameras. The working principle is fundamentally different from frame-based cameras. Mimicking visual processing of mammal eyes every pixel on the chip operates independently from all others and generates an *event* every time it perceives a change in luminance exceeding a certain threshold. The visual information is encoded as stream of single pixel events that have been asynchronously generated. This is fundamentally different from classical camera systems which periodically create frames with a fixed frequency by sampling all pixels at once. In contrast to regular cameras, there is no notion of single pictures or frames in DVS event streams. This reduces the data rate because the event stream is largely free of redundant information and transfers only the relevant information about the observed scene. In a static setting the data stream will only convey information about moving objects; while, when the DVS is moved, most of the events will be generated at object edges or sharp textures changes. In order to be displayable, a time interval of the event stream has to accumulated to form a frame (Fig. 1).

As events are generated asynchronously and immediately after a brightness change, it is furthermore possible to achieve temporal resolutions on the order of microseconds (compared to milliseconds for frame-based systems) which results in strongly reduced latency of information transmission.

The current prototype (Fig. 2) consists of two DVS with fixed geometry, a resolution of 128x128 pixels each and lenses with focal length of 6mm. Their synchronized vision stream is transmitted to a computing stick on which the depth extraction is performed and events are translated into virtual spatial sounds. The sound signal is sent to the users ear via headphones connected to a USB sound adapter.

The two DVS sensors are equipped with an additional microcontroller and require $\leq$0.3W each (80mA at 3.3V. The processing is currently done on an off-the-shelf compute stick which consumes approximately 10W (2A at 5V). We are working on porting the processing part onto a microcontroller which operates at $\leq$ 0.5W (150mA at 3V) and would allow the whole system to operate at $\approx$1W.

With a state-of-the-art cell phone battery we aim to achieve battery lifetimes of around 10h when the processing is done on a microcontroller.

### B. Visual data processing

In a normal use case the DVS produce on the order of 100 events per millisecond, far more signals than the human auditory system can process [5]. We, therefore, downsample the number of events that will be transmitted to the user by a factor of 1000 to roughly 100 per second. In contrast to prior
work, we do not apply clustering strategies as clustering did not prove to increase performance measurably. The distance to the object that generated the event is then extracted using the stereo information of the two DVS stream.

As experiments conducted with an earlier version of this device have shown poor performance of elevation estimation using virtual spatial sounds we developed visual processing as well as sound generation: the user is now provided with a vertical reference point within their field of view. The field of view has been divided into three horizontal stripes of different width. Events are grouped with respect to their specific section. We call the middle stripe focus area. Inspired by the fovea centralis, where human vision is sharpest, we choose to convey a larger amount of information from this area to the user and pick 60% of the sonified events from the focus area. The focus height was chosen to be 4 pixels, corresponding to about 1.5° in the field of view; this matches approximately the minimum audible angle for horizontally separated sounds, i.e. the minimum angle two sound sources must be apart to be perceivable as separate, so that vertical and horizontal resolution become similar. Events from the focal area are furthermore associated with a sound that differs from the sound associated to events from off-focus areas to make them easily distinguishable.

C. Acoustic information transmission

As soon as the position of an event in space is known and it has been selected for sonification it will be translated into a virtual spatial sound, i.e. into a sound whose source appears to lie at the position of the event. This encoding ensures an intuitive presentation of the information to avoid the necessity of long training times before being able to use the device.

Generating sounds that are perceived as coming from an external source in space is a complex task. It requires the modulation of sounds via a head-related transfer function (HRTF) that describes how sounds are changed by the perceivers’ body, head and pinna before they reach the ear canal. As HRTFs depend on the hearer’s anatomy, it is a highly individual function which requires intricate measurements.

To address the need of individual HRTFs for sound elevation resolution, a method for easier personalization of HRTF functions from HRTF databases [8] using body measurements of test subjects was developed. It is a trade-off between accuracy and feasibility and achieved a significant improvement in user performance when estimating elevation of sound source.

Two types of clearly distinguishable sounds are implemented to encode events. A 440 Hz tone for events originating in the focus area and a click sound for events from the off-focal area. They are passed through the head-related transfer function which transforms them into binaural sounds that appear to come from the location where the event has been generated when played to the user’s ear using stereo headphones.

The total latency of the system is approximately 50ms which is mostly due to the way sounds are aggregated before they are replayed. We aim to achieve a latency of 10ms by implementing more sophisticated sound replay algorithms.

III. Tests

In order to test the new processing algorithm, two types of experiments were carried out. The first is based on previous tests [6] to obtain directly comparable results and consists of object detection, size discrimination, and object localization. The second experiment is a common test for visual acuity using the Landolt C based on free online resources [9]. The visual and acoustic processing during the experiments was done using a laptop which provided better controllability and visualization.

The first test was carried out by 11 subjects (9 male, 2 female, average age $\mu = 28 \pm 5$ years). They were seated in 60 cm distance from a white wall on which circles (black on white paper) of different sizes were fixated. We modified the test in [6] to not use screens to display the circles because many artifact events were generated due to the refresh rate of standard monitors. The subjects wore the device positioned in a way that its field of view was centered in the middle of the test area when they looked straight ahead. For each category (object detection (OD), size discrimination (SD), object localization (OL)) twenty trials were conducted where the truth was randomly, equally distributed. Subjects were given assistance to realign the field of view with the middle of the test area after each trial if required.

The second experiment to test the efficacy of the introduced focal area for vertical resolution consisted of the Landolt-C Test where the task is to discriminate between two possible orientations of an optotype (i.e. a standardized letter for acuity tests), here the letter ‘C’ (oriented with opening facing right or left). The vertical visual acuity (vVA) is determined by the size of the letter for which a certain percentage of correctly identified orientations is exceeded. The test group consisted of 5 subjects (3 male, 2 female, average age $\mu = 27 \pm 5$ years). Subjects were again seated in front of a white wall on which a printed optotype ‘C’ was fixated. Due to the fixed absolute size of the printed optotype the distance to the wall was varied to
change the relative size. On average the test lasted 30 minutes ($\sigma = 10$ minutes).

IV. RESULTS

For the first test, we split evaluation of object localization in horizontal position (left, right) and vertical position (up, down) to assess the effects of the new information processing and conveying strategies.

The change in performance for object detection, size discrimination and horizontal object localization is only insignificant compared to the previous version as it was already at a very high level (object detection 99%±1%, size discrimination 96%±5.3%, horizontal localization 90%±8.5%, Fig. 3). An interesting observation is that subjects required more time for the size discrimination than before. The reason behind this is not clear yet and needs to be further investigated (Fig. 4). The introduction of a focal area in the system alongside the personalization of the HRFTs has, however, helped to overcome the inability to resolve the elevation of the artificial sounds and raised the performance from about chance level to a success rate of approximately 95% (Fig.3). In contrast to the prior design subjects explore their environment by moving the field of view by moving their heads, while in prior tests they were asked to keep their heads still.

The vOICe (II) was tested with two different groups of subjects: the first test int Table I (I) was performed by sighted subjects and consisted of a Snellen-E test [5], the second test (II) by blind users with 50-100h training performing a Snellen-E test [10], the phosphene prosthesis used a Landolt-C test with 8 orientations [11], BrainPort was tested with untrained, mostly visually impaired users [12].

TABLE I

<table>
<thead>
<tr>
<th>ETA</th>
<th>Acuity</th>
<th>logMAR Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>This device</td>
<td>$\frac{20}{4500}$</td>
<td>$\frac{20}{5150}$ for vVA $\frac{20}{5500}$</td>
</tr>
<tr>
<td>The vOICe (I)</td>
<td>$\frac{20}{550}$</td>
<td>$\frac{20}{550}$</td>
</tr>
<tr>
<td>The vOICe (II)</td>
<td>$\frac{20}{550}$</td>
<td>$\frac{20}{600}$</td>
</tr>
<tr>
<td>Phosphene Prosthesis</td>
<td>$\frac{20}{550}$</td>
<td>$\frac{20}{600}$</td>
</tr>
<tr>
<td>BrainPort</td>
<td>$\frac{20}{550}$</td>
<td>$\frac{20}{550}$</td>
</tr>
</tbody>
</table>

Vertical acuity for our device; the other devices were tested without discriminating directions. Furthermore, the results were not gained using an identical standardized test, so they cannot be directly be compared. For more details see text.

V. CONCLUSION

We have presented an improved version of a light-weight device that processes visual information from the environment and presents it to a visually impaired user via 3D spatial sounds. Tests showed that former problems with elevation resolution could be resolved with the introduction of a visual focus area and improved head-related transfer functions. However, these tests did not fully cover the field of application as subjects were static and did not have to avoid objects. Real-world scenarios are far more complicated, so further tests with moving subjects need to be carried out. Future work will also be to investigate how performance changes with training, but average test duration shows that the device can already be intuitively and successfully used by inexperienced, untrained persons.

REFERENCES


