Publihsed in Journal of Assistive Technologies, 2014, vol. 8, no. 2, pp. 77-94, which should be cited to refer ot this work.

DOI: 10.1108/JAT-08-2013-0025

See ColOr: An extended Sensory Substitution Device.

Audio-based Sensory Substitution Devices (SSDs) perform adequately when sensing and mapping low-level visual features into sound. Yet, their limitations become apparent when it comes to represent high-level or conceptual information involved in vision. We introduce *See ColOr* as an SSD that senses color and depth to convert them into musical instrument sounds. In addition and unlike any other approach, our SSD extends beyond a sensing prototype, by integrating computer vision methods to produce reliable knowledge about the physical world (effortlessly for the user). Experiments reported in this article reveal that our See ColOr SSD is learnable, functional, and provides easy interaction. In moderate time, participants were able to grasp visual information from the environment out of which they could derive: spatial awareness, ability to find someone, location of daily objects, and skill to walk safely avoiding obstacles. Our encouraging results open a door towards autonomous mobility of the blind.

INTRODUCTION

Multisensory perception theories state that sensory inputs in our brain are never processed separately [1], [2]. On the contrary, information from different sensory modalities is integrated by the nervous system enabling coherent perception of the world, which ultimately leads to meaningful perceptual experiences [3]. As a consequence, for example, what we see, somehow and somewhat, is always influenced by what we hear, and vice versa. A remarkable example to this is the McGurk effect [4] that shows experimentally how vision overwrites hearing. In light of these assumptions, it is plausible to think that a particular sensory perception could be elicited through a sensory pathway that is not typically responsible for it. In fact, this is the idea central to neurological behavior that neuroscientists have termed cross-modal transfer [19], [21]. They state that due to all this sensorial interconnection in our brain, visual-like experiences may be attained through senses others than vision [19], [21].

Sensory substitution then, refers to the mapping of stimuli of one sensory modality into another. This is usually done with the aim of bypassing a defective sense, so that associated stimuli may flow through a functioning sense [5]. In general, sensory substitution argues that when individuals go blind or deaf, they do not actually lose the ability of seeing or hearing, rather they become incapable to convey external stimuli to the brain. When the working of the brain is not affected, in most of the cases a person who lost the ability to retrieve data from their eyes could still create subjective images by using data conveyed from other sensory modalities (e.g. auditory pathway) [5].

Sensory substitution may be achieved using invasive methods that collect external signals and transduce them into electrical impulses for the brain to interpret them naturally [5], [23]. Thus, stimulation of the brain without intact sensory organs to relay the information is possible [5], [6], [7]. However, in this work, we address non-invasive sensory substitution devices also known as SSDs. These devices use computational interfaces to transmit sensory information (of a substituted sense) captured by an artificial modality (artificial sensor) to another human sense (substituting sense) [23]. In other words, they translate sensory information in order to enable perception through another one than the originally responsible sense [8]. This idea largely relies upon the concept of brain plasticity [7] that denotes the self-adaptation ability of the brain to the deterioration (or absence) of a sense [7]. This is a natural mechanism that allows people devoid of a sense to somewhat adapt and compensate through other sensory pathways. For instance, cortical re-mapping or reorganization happens when the brain is subject to some type of deterioration or injury [9].

Motivation

Vision is a phenomenon that entails both sensation and perception [10]. Sensation is the low-level -biochemical and neurological- feeling of external visual information as it is registered (sensed) by the eyes. The visual sensation alone does not imply the coherent conception (or understanding) of external visual objects [10], [11]. Therefore, following a sensation, perception appears as the mental process that decodes the sensory input (sensation) to create awareness or understanding of the real-world [10], [11], [12]. In short, we perceive the world through sensations, though we derive sense from it (vision comes into being) only when perception takes place [10], [13]. In this work, we argue that current SSDs have been intended to provide a substitute to sensation, while the perceptual experience has been left mostly unattended. The underlying problem is that the human visual system is known to be capable of 4.3×10^6 bits per second (bps) bandwidth [14], [15]. Yet, senses intended as substitutes can hardly reach 10^4 bps at most (i.e. hearing) [14], [15]. In this light, even though a cross-modal transfer may apply, it is hard for mapping systems to overcome the large sensory mismatch between vision and other sensory pathways: *if hearing does not even provide room enough to convey visual sensations; actual visual perceptions are therefore very unlikely*.

Importantly though, we do not think of visual perception being unattainable through long-term use of current SSDs. Simply, it implies a tough/long learning process that in any case, will yield inaccurate approximations of vision, if at all. Further, we argue that any visual perception we can achieve through hearing will always need to be reinforced, in order for the substitution to be: practical, fast, accurate, and let users act as though they were actually seeing [18]. Visual perception is a holistic phenomenon that emerges from complex information unlikely to be encoded into hearing (shapes, perspectives, color, position, distance, texture, luminance, concepts etc.) [10], [11], [12], [3], [16], [17]. In fact, 'normal' vision is itself constrained by top-down knowledge that produces the kind of information

that sighted individuals typically achieve without conscious effort [12], [13], [18]. Our hypothesis is that nowadays, all this amount of data cannot be supplied efficiently in SSDs, unless we integrate more advanced methods that lie beyond mere visual-to-audio mapping (e.g. computer or artificial vision techniques). In this spirit, we shall not abandon the encoding of low-level features into sound for sensory substitution. Rather, we would like to augment such an approach with the use of computer vision and image processing to deal with high-level information that usually surpasses the bandwidth of audio. Whether this strategy will lead us to an SSD more learnable, practical, easy to interact with, and chiefly functional? This is the fundamental research question underlying this work.

1. STATE OF THE ART

Back in the late 1960s, Paul Bach-y-Rita introduced the first SSD [22] prototype known as TVSS (Tactile Visual Sensory Substitution): black-and-white camera images, instead of going to a TV screen, were sent to a matrix of vibrators in contact with the skin of the back of a blind participant [5]. However, first attempts to substitute vision using the auditory pathway came later on. In 1975, Raymond M. Fish [24] used a sequence of tone bursts to represent image pixels: vertical locations were determined by the frequency of the tone, whereas horizontal positions were represented by the ratio of sound amplitudes presented to each ear of the blind user. Following in 1977, T.Bower writing for the Newscientist magazine [25], reported on several new devices that provide blind babies with sound information about their environment using ultrasonic echolocation similar to that of a bat. In fact, in the 1980s there was a proliferation of these ultrasound-based devices with improved features, known as ultrasonic pathfinders [26], [27]. One of the most popular examples is “The Sonic Guide” [26] that constantly irradiates from objects into audible sound to reveal distances.

In 1992, the vOICe [28] was introduced as a SSD to sonify 2D gray scale images, allowing its user to “see with sound”. The sonification algorithm presented in this work uses three pixel parameters: horizontal position, vertical position and brightness to synthesize a complex audio output describing the whole image (soundscape). Later in 1997, Capelle *et al.* [29] proposed the implementation of a crude model of the primary visual system [29]. The auditory representation of an image was similar to that used in “The vOICe” [28] with distinct sinusoidal waves for each pixel. Gonzalez-Mora *et al.* [30] developed in 1999 a prototype using the spatialization of sound in the three dimensional space [30]. The sound was perceived as coming from somewhere in front of the user by means of head related transfer functions (HRTFs). In 2000, Soundview [31] represented a single point of color as sound scaled activated by the velocity of haptic exploration of an image on a tablet with no haptic feedback (vibration, temperature etc.). *TheVIBE* [32], was later introduced in 2008 as a visual-auditory substitution system that converts video-streaming into auditory streaming. The sound generated is a weighted summation of sinusoidal sounds produced by virtual "sources", corresponding each to a "receptive field" in the image [32].

More recently in 2009, the Kromophone [30] and The Shelfscanning [31] were introduced. The former takes the input from a webcam and chooses the center pixel and maps its color into several superimposed sounds [30]. As for the latter, it was intended to empower visually impaired individuals to shop at their own convenience from a shopping list using computer vision methods such as object recognition, sign reading and text to speech encoding [31]. Then in 2010, Michał Bujacz [32] presented an algorithm for sonification of 3D scenes including the use of segmented 3D scene images, personalized spatial audio and musical sound codes. Thus, virtual sound sources originating from various scene elements were generated [32]. In 2012, The EyeMusic [19] emerged as a tool that provides visual information through a musical auditory experience: in about 2 seconds. Colors in a 24x40 pixel image are first segregated into red, green, blue, yellow, white, or black. Then each color is encoded through a timbre (e.g. white = piano, blue = marimba) [19]. Also in 2012, Rebeiro *et al*. [33] coined the term auditory augmented reality in reference to the sonification of objects that do not intrinsically produce sound, with the purpose of revealing their location to the blind. They use spatialized (3D) audio to place acoustic virtual objects that create the illusion of being originating from real-world coordinates.

It becomes apparent in the literature that experiments on mobility/exploration using SSDs are rather few. Their usability to substitute actual vision, therefore, remains largely uncertain in practice. Most of the cited works are intended to translate static images into sounds. For instance, the vOICe [28], one of the most popular SSDs, takes two seconds to encode a 100x100 image, resulting in a complex sound that requires also time to be fully understood. This makes it hard for it to function as a real time mobility tool. Mobility encompasses three main tasks: understanding the near space global geometry; avoiding obstacles; and focusing on a specific goal to reach, for instance looking for a specific door, a person, etc. Moreover, in contrast with the vast majority of works, a key aspect of See ColOr is to provide a substitute to color for the blind users. Likewise, approaches that account for depth are still few. To the best of our knowledge, there is no approach other than ours, attempting to code both color and depth into sound.

More importantly, despite the increased use of computer vision nowadays, surprisingly its use in SSDs is rather vapid. This is curious since one would think that devices meant to substitute natural vision should be heavily based on artificial vision. One is left with the thought that it seems unpractical to have methods to enable computers to “see”, and not to target it to the benefit of blind in combination with SSDs. In our view, there should be, at least, a marked tendency to the use of robust and stable computer vision technologies to strengthen the weakness of existing electronic aid device.

2. See ColOr: Seeing Colors with an Orchestra.

See ColOr is a multimode aid (SSD + computer vision) engineered with the following components (Figure 1): 3D camera (ASUS Xtion PRO LIVE); Bone-phones (AfterShokz); and 14" laptop. Occasionally, a tactile tablet (iPad) is used to enhance interaction. Users can switch between a local and a global module as they explore the nearby environment. In any case, an alerting system and a recognition module are always running in the background to prevent the user from hitting obstacles, and to inform about the presence of known persons/objects (Figure 5).

Figure1: A blind individual wearing See ColOr (photograph with permission). Note that the tactile tablet was not subject of study in this paper. Also, the ears of the user remain uncovered due to the bone-phones.

2.1 The local module

The local module provides the user with the auditory representation of a row containing 25 points of the central part of captured image [35] (Figure 3). These points are coded into left-right spatialized musical instrument sounds, in order to represent and emphasize the color and location of visual entities in the environment [37], [38], [39]. The key idea is to represent a pixel as a sound source located at a particular azimuth angle [38]. Moreover, each emitted sound is assigned to a musical instrument and to a sound duration, based on the color and the depth of the pixel, respectively (Figure 2). The local module allows the user to explore a scene as he can move the head to scan it. However, since the perception of the user is focused only on a small portion, peripheral vision and global perception become unattainable. Having access to more than one portion of the image simultaneously would bring many advantages. For instance, it would be possible to compare several distant points. Note that such a task is unachievable using this local module. To rectify for this deficiency, we introduced the global perception module that allows the user to explore several points with the fingers.

Figure 2: Empirical association of colors and instruments sounds in See CoLOr: Oboe for red; viola for orange; pizzicato violin for yellow; flute for green; Trumpet for cyan; piano for blue; and saxophone for purple. The duration of the sound is defined by the distance of the pixel: the closer the pixel the shorter the sound. The distance of the pixel is given by the 3D camera.

2.2 The Global module

In the global module the image is made accessible on a tactile tablet that makes it possible for the user to compare different points and explore the scene in a broader context [35], [40] (Figure 3). A user may rapidly scan an image sliding one or more fingers on this tablet. The finger movements are intended to mimic those of the eyes. A finger tap on the tablet activates a sound that codes the color and depth of the corresponding pixel (Figure 2). The spatial position of this pixel (left to right) is mapped to the hearing by directional virtual sound sources (spatialized sound) [36], [37], [38], [39]. In theory, the entire image resolution (460x640 using the Kinect camera) is made accessible to make the most of the camera information (Figure 3). However, only as many points as fingers can be accessed simultaneously not to reach the limits of the audio bandwidth [36]. By and large, the global module is intended to promote a more proactive interaction to selectively explore, to discover points of interest, make comparisons, and, in general, enjoy a greater sense of independence.

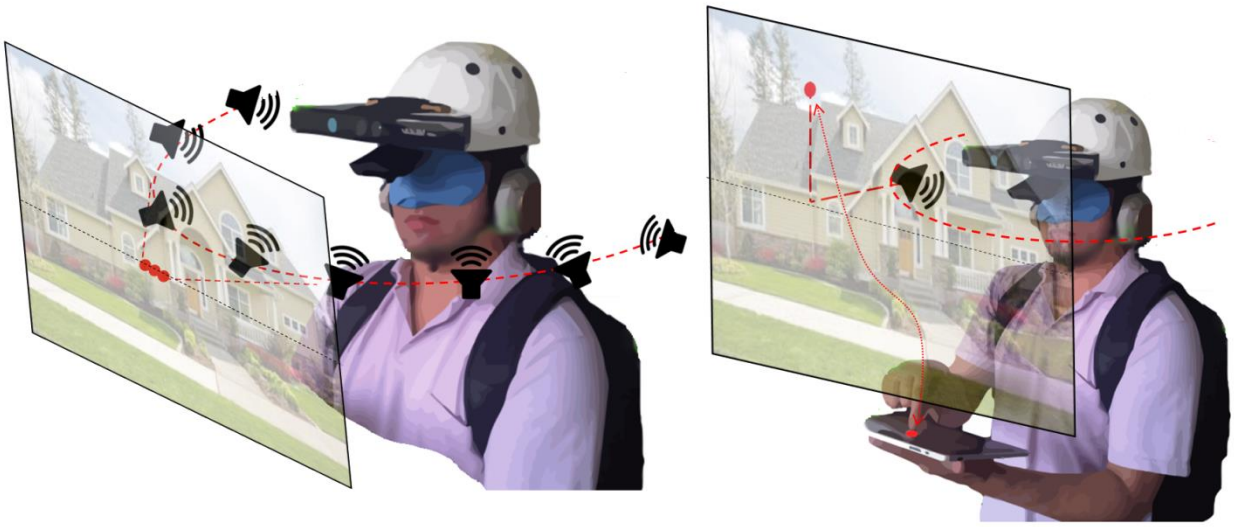

Figure 3: To left of this figure the sonification with the local module is illustrated. There are 25 pixels and 25 sources in this module. To effects of visualization however, only 3 and 8 are respectively displayed. Note that when the row of 25 pixels (points) related to the central part of the image is mapped into sound; it is also augmented to cover the whole azimuth-frontal auditory field. To the right of this figure an illustration of the sonification with the global module is presented. Now, only the pixel tapped with the fingertip is sonified. Note that the use of spatialized sounds gives the user awareness of the lateral position of the point (from left to right). In this illustration, therefore, the source matches the position of the point horizontally, but not in elevation. It is well known that rendering elevation is much more complicated than lateralization [37].

2.3 Toward higher levels of vision:

2.3.1 The alerting module

Since the functionalities just mentioned are limited to describe local portions using low-level features such as color and depth, they might fail to reveal cognitive aspects which often determine regions of interest within a picture [33]. The purpose of alerting system is to warn the user whenever a hazardous situation arises from obstacles lying ahead [35]. Once the system launches a warning, the user is expected to suspend the navigation not to bump into an obstacle. This allows the blind persons finding a safe, clear path to advance through [40], [41]. Roughly, when potential obstacle in the video presenting a distance below one meter continues to approach over a given number of frames, the user must be alerted (Figure 4, right). We use image processing methods to this end. It is worth noticing that the alerting system is an autonomous algorithm that demands no user intervention and runs in parallel to the rest of the modules. Thus, users will focus on the exploration without loss of safety (Figure 4).

Figure 4: A three-frame sequence of a real video as a visual example of how the alerting system and the recognition module work together. By default, the recognition module detects known elements nearby (frame one and two) and notifies the user. When an element within the scene (known or unknown) becomes potentially dangerous (i.e. user is approaching and likely to bump into it) the alerting system activates an alarm. Notice that the information displayed in this figure must be presented to the user via audio (See ColOr's goal).

2.3.2 The recognition module

See ColOr also uses computer vision techniques to process higher visual features of the images in order to produce acoustic virtual objects (Figure 4). Actually, we recognize and then sonify objects that do not intrinsically produce sound, with the purpose of revealing their nature and location. The recognition module is a detecting-and-tracking hybrid method [44] for learning the appearance of natural objects in unconstrained video streams [42],[43],[44], [45]. Firstly, there is a training phase to learn the distinct appearance of an object of interest (scale, rotation, perspective, etc.) [44], [46]. This is an off-line process that must be carried out by sighted individuals. Afterwards, a visually impaired user that performs exploration with See ColOr is informed about the presence of learned objects in real time, if any (Figure 4, left, center). Overall, this module allows the blind noticing serendipitous discoveries, seeking a specific target, as well as avoiding obstacles [35] (Figure 5).

Figure 5: Overall operation of See ColOr. While the user explores the environment, he is aware of the entities surrounding and alerted of potential obstacles. Notice that the mental modeling of a chair becomes easier when the user already knows "this is a chair". Further details such as color and shape later come with the exploration itself (either using the local or the global module).

3. EXPERIMENTS

3.1 Detecting a colored target

This study concerns the capacity of blind users to perceive, through the audio feedback, salient points (colored target) in the environment. This allows the capture of relevant events and changes in properties of the world. Further, this study evaluates the ability of the user to interpret the mapping between spatial relations into sound. Locating something nearby is somehow a constantly happening task that we carry out in pursuit of several goals such as: reaching somewhere, avoiding crashes, making ourselves aware of the environment layout, and so forth. In short, "where is something" is a key query that supports in great extent, nature exploration and navigation of scenarios.

Hypothesis: Independently of the spatial location of a target on the azimuth plane, a blind individual can reach it safely, with minimum effort, and in moderate time (Figure 6).

Procedure: Firstly, the concept of See ColOr was explained to the ten participants. Next, they wore the system and interacted with it freely for ten minutes. The system was adapted to react only to the red color and remain quiet otherwise. This eases the localization process and avoids the memorization of our entire sonic code. They were asked if they understood the working of the system and they finally consented to taking part in the experiment.

Experiment: To obtain quantifiable and repeatable results, we modeled and limited this experiment as follow: a red target is placed at certain location with respect to a blind participant (sitting on a spinning chair) within a radio of 2 meters. The participant is then asked to detect the target and walk towards it (Figure 6). The test consisted of 5 trials for which the target was located at the following five positions, though not necessarily in this order: in front, to the right, behind, to the left and randomly with respect to the participant. For the latter, we spin the chair so the direction the user ends up facing is unknown. To achieve his goal, the participant is asked to spin the chair looking for the target that emits the sound of red. This emitted sound might come from either right or left, depending on the alignment of the user with the target as he slowly spins the chair (i.e. sound spatialization). Once the target is thought to be centered, the participant is asked to stand up and walk to reach it. In this experiment the alerting system is activated, meaning that while spatialized sound leads the user to locate the target, the alerting system prevents him from bumping into it.

Figure 6: The spinning chair experiment (Detecting a colored target).

Our goal is to measure the impact of the independent variable "Target Position" *TP* ($TP \in [0°, 360°]$) on the variables time (t) and crash (c). These variables are descriptors of the efficiency on reaching the goal. For instance, a blind person using no aids (or even using a cane) has no other option than guessing the location of the target. We want to prove that See ColOr makes it possible to reach this goal in stable moderate time, while keeping safety.

3.2 Awareness of walls

This study concerns the ability to be aware of oneself in space. In this experiment we evaluate how efficiently See ColOr provides awareness of spatial relations to the blind (i.e. perceive an entity in relation to oneself). Spatial awareness let a person be included into space, causing understanding of his location and the location of objects in relation to his body. In grasping these relationships, persons come to mechanize concepts such as distance and location. For example, a person with spatial awareness understands that as (s)he walks towards a door, the door is becoming closer to his/her own body. This understanding is all

achieved during our earliest age. Unsighted people, nonetheless, lack this ability and their positioning in relation to the world is a thorough trial-and-error process.

Hypothesis: By means of his auditory pathway, a blind person can regain awareness of entities lying on his path (Figure 7).

Procedure: The participant is taught two strategies that lead the goal, and then he is given 10 minutes to practice each:

- Assessing the sound emitted by the wall itself: The closer the wall, the faster the rhythm of the sound.
- Relying upon the alerting system: The participant progresses without heeding the rhythm but, waiting for a warning to be launched.

Experiment: A participant is tasked to walk certain distance straight towards a wall. The distances evaluated in this experiment were 3, 5 and 8 meters. The goal here is to be aware of the distance to the wall accurately enough to stop timely and not to collide; yet close enough to reach out and touch it (Figure 7). The goal is to be scored as follow:

- fully achieved (i.e. participant could stop few steps ahead, and managed to touch the wall);
- partially achieved (i.e. participant could stop but, failed to touch);
- failed (i.e. participant required our aid not to bump into the wall).

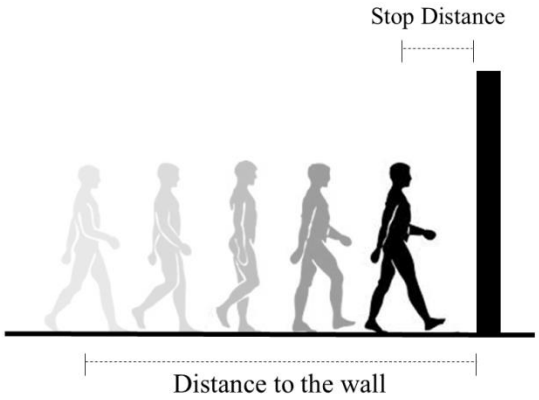

Figure 7: Awareness of a wall experiment.

When it comes to sensing a wall, blind people successfully rely on a white cane. However, our aim is to show that regardless the distance being walked, See ColOr might be as reliable as the cane. If this occurs, See ColOr will additionally prevent users from having their hands occupied and liberate mental attention. As for this latter, this experiment was conducted with and without an alerting system doing the detection automatically. We

would like to see whether or not removing the user attention from this task has some negative effect on the overall performance.

3.3 Finding a person

Based on surveys with visually impaired and blind users, the authors in [33] claim that face detection and recognition were suggested as highly desirable features for an assistive device. For this reason among others, in this study we use face recognition as a mean to verifying location and identity of people within the environment. We want to assess the effectiveness of See Color in guiding the route that leads a blind individual to meet someone nearby. Particularly, for blind individuals, ignoring information about approaching people generally represents a missed opportunity to socializing. Visual cues revealing distinguishable features are imperative for achieving the recognition of a face (or person). Nonetheless, all this large amount visual information can be condensed into audio cues. See ColOr's final aim is to achieve automatic labeling of persons (stored in database) so as to reveal their identities to the blind upon encountering them; quite like the visual system.

Hypothesis: Relying upon none of the typical strategies such as sensing noise, movement or voices, a blind individual can still find, recognize, approach and shake the hand of a person standing within a controlled environment, far more efficiently that he does by chance. (Figure 8)

Procedure: This experiment requires no training, for the detection is done automatically. Prior to the experiment, however, we learn our See ColOr to identify the face of a particular person (target). Once target is detected the system announces his presence by speech (i.e. "person").

Experiment: The blind subject is placed at the center of a 20x20 squared meters room. He is provided with the system and asked to scan the environment (360°) until the speech alert-detection is launched. To avoid any collision, we keep the alerting system on, so the participant can walk safely. The reasons for the participant to move might be twofold: either he found the target and wants to approach, or the target is yet undetectable due to the distance. Once the target is focused, the participant has to advance towards his position to shake his hand, never bumping into him or into a wall (Figure 8).

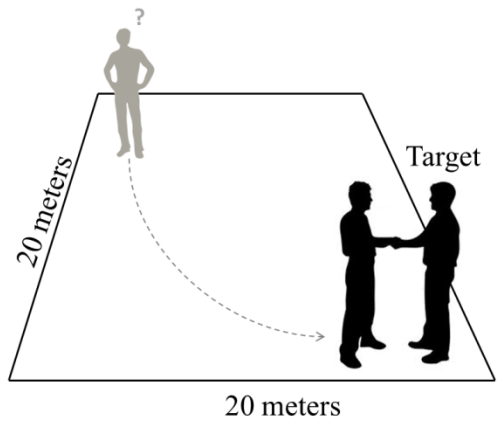

Figure 8: Finding persons experiment.

The target person stands still at each corner of the room for trial (4 trials). The goal here is to show that See ColOr reaches the behavioral criterion of visual substitution [18]. This criterion establishes that if a person can carry out normally some functions ascribed to vision, the substitution indeed resembles vision. Without See ColOr, a blind individual has to rely on tactile inspection of the walls, which is mostly uncomfortable. Worse, since the participant does not know that the target is leaning on a corner, the task becomes a trial-and-error exercise that lasts in average 9.3 minutes[1]. And, unless accompanied by a cane, the participant is likely to bump into the target.

3.4 Grasping objects

This study concerns the retrieving of daily objects. Here we look forward to evaluating the ease with which See ColOr orients a user to locate and seize small items. For instance, when sighted individuals drop something, retrieving it is quite an easy task; by contrast blind people struggle to get into do it. In fact, fallen objects very often yield embarrassing situations that might lower their feeling of dignity. Further, it is particularly useful to allow the visually impaired reaching daily objects they otherwise could fail to notice.

Hypothesis: Using no tactile inspection, blind people can locate/identify small daily objects so accurately as to reach them. (Figure 9).

Procedure: This experiment requires no training, for the detection is done automatically. Prior to the experiment, however, we learn our See ColOr to identify six different objects, namely: sunglasses, keys, glass, cap, remote control, and a landline telephone. Objects are identified by their name when detected.

Experiment: The six elements are arranged on a 2.5x0.75 m table. We put the blind participants to the test by asking them to find one of the items, with the aid of See ColOr. The participant has to scan the table walking back and forth, while paying attention to the

audio cues emitted by the items. To this end, they are allowed either use the cane or tap the edge of the table. Yet, the participant by no means can touch the objects (Figure 9). The participant has to seize the asked object with the hands and continue to seek a new one until he collects the six of them. The order in which elements are removed from the table is chosen randomly and changes among participants. This lets us evaluate also the efficiency of our system in finding a particular object in different environments.

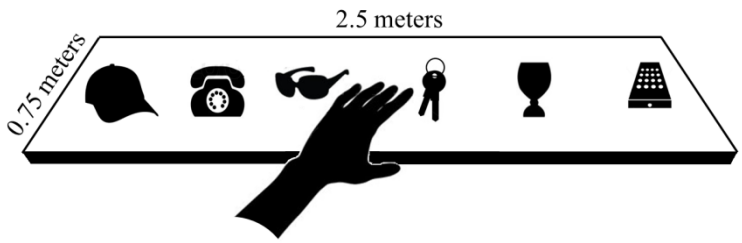

Figure 9: Grasping objects experiment.

4. EXPERIMENTS RESULTS

Using See ColOr, a participant was able to detect, center and reach a surrounding colored target quickly (2.35 minutes in average). Figure 10 shows the time distribution for ten participants undergoing five different trials (random, 180°, 270°, 90° and 0°). The mean time (in red) of the trials showed no significant correlation with time (t), i.e. p-values >0.05. In other words, the variable 'Target Position' ($TP \in [0°, 360°]$) had not significant effect on the time t. By and large, the observation was made that the search of the target presented no difficulty in this experiment, as the participants were confident to the use of the spinning chair.

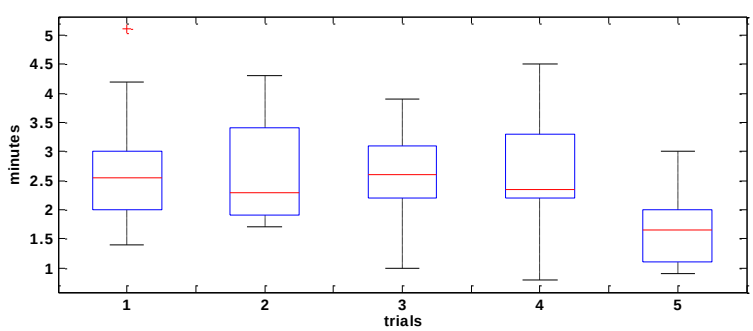

Figure 10: Results first experiment "Detecting a colored target". Trials correspond to random, 90°, 180°, 270°, and 0° location of the target respectively.

Consequently, out of 2.35 minutes that a participant spent (in average) in this test, 1.7 minutes (70%) were used up in walking towards the target. Therefore, it took only 0.65 minutes (30%) to encounter the target, but most of all, to center it (i.e. to master the spatialized sound). This fact is reflected on Figure N that reaches a minimum mean time when the target had no need to be centered (it was located in front, *TP*=0°). Also, the fact that participants needed 1.7 minutes to walk 2 meters (radio), reveals their misgiving to abandon the cane as they fear to collide. Importantly though, we need to mention that collisions (*c*) never happened. Moreover, comparisons of See ColOr users against participants with/without a cane in this experiment were worthless and even dangerous. Blind individuals not wearing See ColOr but a cane or nothing, had to guess the location of the target. Therefore, the likelihood of success is negligible, the time ends up depending on chance and collisions are likely. Overall, this experiment concludes that reaching a target (with See ColOr) is not randomly accomplished. In fact, it can be performed in moderate time, safely and effortlessly.

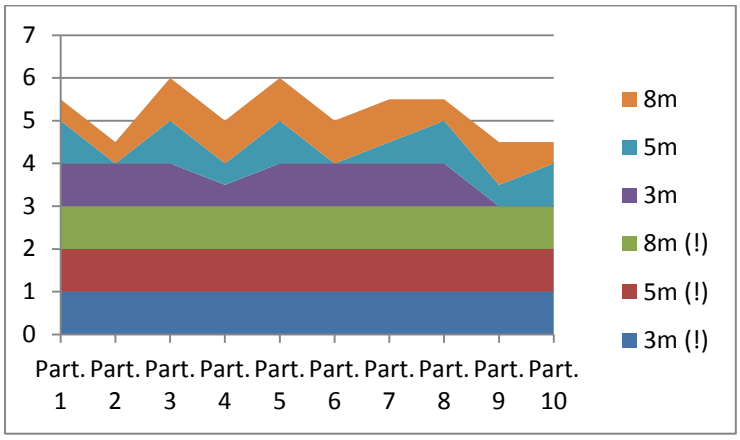

Figure 11: Results first experiment "Awareness of a wall" from 8, 5 and 3 meters. With (!) and without an alerting system on.

The chart showed in Figure 11 reports on the results attained in the experiment 'Awareness of a wall'. In this chart the performance of participants were scored as follow: 1 for a goal fully achieved; 0.5 for a goal partially achieved; and finally 0 for a goal unachieved. Notice that tests marked as (!) indicate they were carried out with the alerting system on. Likewise, tests not marked were performed while the alerting system was deactivated. Even though those tests accomplished with no the alerting system were not bad at all (only 4 failures out of 30 trials). The use of the alerting system is strongly advisable, as it leaves no room for failures. Only two participants (3 and 5) got the maximum score (6), showing no need to use the alerting system. This is quite a good example to show how the automation of visual tasks in SSDs, leads to enhanced performances. This experiment reveals that (using our extended SSD) blind individuals can sense walls almost unconsciously, pretty much as a sighted person. Notice that such an advantage serves to free the auditory bandwidth, which is highly convenient in audio-based

SSDs. For the visually impaired; remaining perceptual capacities are further lessened by the focus needed for the mobility and orientation tasks. In fact, the alerting system brings a greater sense of independence since users need no longer focus on sensing unexpected obstacles. Eventually, this could serve as an alternative to the use of a cane. Finally, we found no evidence (as it is apparent in Figure) to conclude that the distance in this experiment affects the accuracy at all.

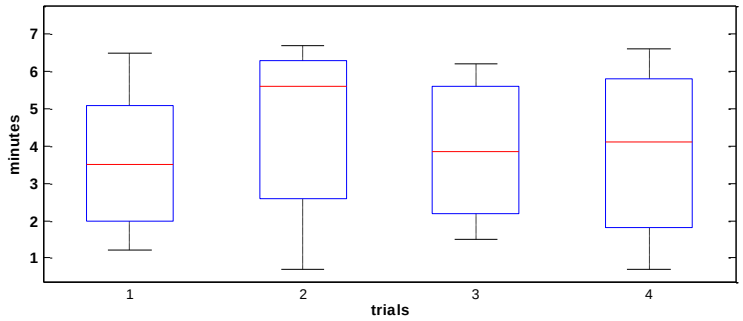

Figure 12: Results first experiment "Finding a person". Trials correspond to each corner of the room.

As for our third experiment, with results revealed in Figure 12, statistical analysis of these data reports that in average a participant takes 4.1 minutes to find someone, approach and shake hands. Compared to users that completed the test without See ColOr in 9.3 minutes (average), our results are fairly efficient. Amid *et al* [19] show that blind individuals using an audio-based SSD (without computer vision) take 70 hours training to start recognizing cartoon-like faces and yet, it yields no mobility aid. We claim that to great extent, See ColOr met the behavioral criterion [18] in this experiment: blind people could behave and carry out functions ascribed to vision with autonomy, safety and efficiency. Overall, with no human aid they could tap the chance to socialize and be aware of the people nearby. Although we had exceptional cases of 0.7 minutes, the first quarter of all the tests conducted fell below 2.05 minutes. And globally speaking, the 75% of the trials never surpassed 5.8 minutes. Last but not least, in any case a participant using See ColOr reached 7 minutes, which is much lower than 9.3 minutes required otherwise.

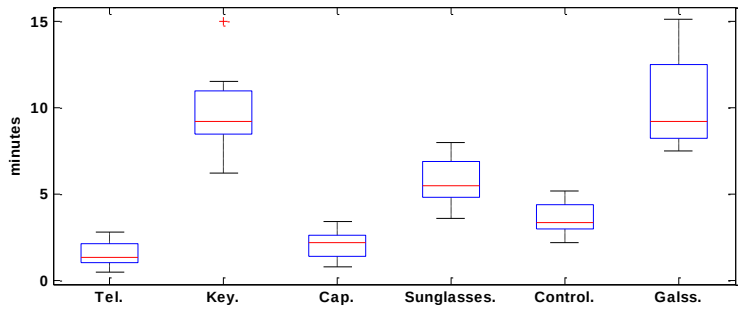

Figure 13: Results first experiment "Grasping objects".

For our last experiment whose results are shown in Figure 13 there was a marked relation between time and size of the sought object: the smaller the object the longer the time. Accordingly, for finding objects such as the keys, the sunglasses and the glass, participants were fairly faster with their hands (tactile feedback) than with See ColOr. This has to do with the recognition rate of the computer vision method that was implemented. The problem underlying is that small items (such as the keys) are devoid of distinguishable features, so detection algorithms are bound to fail. Another issue is that the detections were constrained to the central part of the image only (i.e. 20% of the picture). This was done to give the user a spatial reference with respect to the target (if detected, the target is right in front). Thus, the peripheral views of the camera were missed and hundreds of detections in the lateral parts of the picture ignored. Consequently, the searching becomes a tough trial-and-error process to center an object with unknown location. Accordingly and regarding the speech (of detected objects) in See ColOr, for future experiments we must: spatialize the sound to represent left/right, modify the pitch to represent top/bottom, and adjust the volume to represent the depth of the objects. Also, further work on feature detection and description needs to be done before reaching more consistent results. As for regular objects such as a cap or a telephone the time the use of See ColOr to find them was very convenient. In average, all participants found the telephone and the cap in 1.35 and 2.2 minutes respectively. This is fairly good taking into account that they do not need to use their hands to touch the table all around, which is often uncomfortable for them in public spaces. Last but not least, in this constrained experiment users knew that the objects were located on a table, so the hand-searching was tolerable. In real scenarios however, this is not the case and the use of See ColOr comes more in handy.

5. DISCUSSION AND CONCLUSIONS

Designing electronic SSDs continues to be a difficult task for a number of reasons. While generally promising, there are still significant gaps in our understanding of these technologies. First of all, there is the biological sensory mismatch between visual information and the rest of the senses. For instance the auditory pathway, even though useful for presenting low-level visual features through sounds, is severely limited when tasked with the analysis of multiple sound sources (necessary to model complex visual information). Nevertheless, we have shown that by leveraging the strengths of computer vision methods, we can build a SSD that is capable of condensing visual information and orient the blind efficiently towards purposes, otherwise barely achievable (Figure 14, http://youtu.be/qLu1yh8XVmk).

While we do not abandon the idea of sonic codifications to represent low visual features, our interest is no longer focused on refinement of these codes to model more complex visual

information [28] [47] [48]. To this end, we rather advocate the use of more practical (from the user's point of view) approaches such as computer-vision-based guidance. We do so, because remaining perceptual capacities of the blind are severely lessened due to the complex interaction imposed by current SSDs. This is to say, the interpretation of complex sounds whose lengthy calculation, besides, slows down the interaction. This is the main reason that despite the proliferation of SSDs, many of these devices still lag behind practical aspects.

Figure 14: Random pictures of our experiments. Leftmost: a blind individual finding a red target, while sippining on a chair. Left: a blind individual sensing a wall. Right: A blind person finding someone. Rightmost: A blind participant grasping targeted items.

Of course, total visual substitution is still far from being achieved, regardless the method. Nonetheless, unlike many, See ColOr is a utilitarian prototype (capable of functioning) that substitutes several features of vision at expense of relatively little user effort (Figure 14). Graham *et al* [47] urge researchers to create functional SSDs: "An ideal device would be intuitive to learn, pleasant to listen to, and capture relevant visual information in sufficient detail". In this regard, See ColOr's main idea is to encode colors into sounds (produced by instruments) that are by no means unpleasant. As for capturing relevant visual information in sufficient detail, See ColOr not only captures color and depth but, also reveals cognitive aspects which often determine regions of interest within a picture. Also, in our recognition module the sonification of virtual objects is usually achieved through natural speech, as it is one of the most powerful (and the most used) form of auditory communication [49]. Besides, we advocate the use of natural speech to label objects, against soundscapes, as it prevents users from spending 70 hours of training (an above).

Some may argue that the use of computer vision (and speech) triggers just visual imagery rather than vision via sensory substitution. We respond to this saying that besides being quite unpleasant and extremely difficult to understand, soundscapes might be subjected to the same phenomenon. Ward et al [18], for instance, enquire whether the vOICe's users are likely to tap their prior experience of vision to augment their "visual-like experience". Furthermore, a plausible conjecture is that after 50 hours of training one is more likely to simply develop an associative pattern between a sound and an image, rather than a visual experience as such. In fact, Poirier et al [50] conclude that mental imagery (and not cross-

modality) is predominant in blindfolded sighted subjects or not congenitally blind using SSDs. However, it is fair saying that Amedi *et al*. [7] show that after 70 hours of training in soundscape identification, congenitally blind individuals started to show some activity in their visual cortex. Yet, whether this activity corresponds to actual visual experiences remains as a philosophical debate on consciousness [21], [3], [10], [13], [16]. In the end, if computer vision does not meet the most restricted definition of sensory substitution. We still can use it as a complement to SSDs, as long as it gives rise to more efficient aids for the visually impaired (Figure 14).

While positive acceptance of bone-phones lessened the skepticism in participants reluctant to cover their ears, concerns still linger in regard of the size of our prototype. Participants stressed that besides being functional, an aid system must be wearable and comfortable. Particularly, we highlight here the request to relocate the camera. Over the years, blind individuals lose the instinctive notion of pointing their heads forward. Thus, in many occasions they tend to walk with a head down posture. Suggestions were made about wearing the camera at breast height alternatively. By and large, participants advise against the use of methods other than voice, for labeling objects. Participants broadly showed enthusiasm and were acquiescent for the use of See ColOr (Figure 14). Some rewarding testimonies in this sense can be watched on the web (http://youtu.be/2rNTWTpu1-8). Those who had a guide-dog, however, expressed little interest in swapping it for current technology. In the end, what all this says to us is that while it would be possible to use See ColOr to help visually impaired individuals with color and depth perception, there remain issues that need to be solved first such as, portability and user adaptation. In recognizing this, Neil Harbisson[1] has been quoted as saying: "At the start, though, I had to memorize the names given for each color, so I had to memorize the sound notes, after some short time, all this information became a perception. I did not have to think about the notes. And after some other time, this perception became a feeling. I started to have favorite colors and I started to dream in colors". Thus, we do think that training in central to the use of See ColOr as a visual aid confortable enough to be broadly accepted.

Finally, to make our See ColOr usable in real life situations, our future research must flow in two directions: portability and high-level computer vision. Firstly, we could investigate thoroughly parallel or high-performance computing, with a view to increasing efficiency of our system for embedded commercial prototypes (e.g. based on FPGAs). More importantly, we need to make it work outdoors, which has not been yet attempted due to limitations of current time-of-flight cameras. To this end, our research will focus on stereo-vision technologies that meet both, accuracy and lightweight. As for our future research in computer vision, we are particularly interested in adding to See ColOr action/pose estimation and attribute classification commonly used in high-level computer vision. For

[1] A color blind individual who is well known for wearing a camera that traduces colors into sounds, in nearly the same way as See ColOr does. (https://archive.org/details/NeilHarbisson_2012G)

instance, poselets[2] are intended to recognize gender, hair style and types of clothes in natural scenes and real scenarios. In this way, much relevant information could be made automatically accessible to unsighted individuals.

REFERENCES

[1] Ghose, D &Wallace, M 2012, 'Impact of response duration on multisensory integration', Neurophysiol, vol. 108, no. 9, pp. 2534-2544.

[2] Aaron, N, *Multisensory Research Laboratory at Vanderbilt University*, Laboratory of Mark Wallace. Available: http://www.kc.vanderbilt.edu/multisensory/index.html. [Accessed 1 May 2013].

[3] Koch, C 2012, Consciousness: Confessions of a Romantic Reductionist, Cambridge: MIT press.

[4] McGurk, H & MacDonald, J 1976, 'Hearing lips and seeing voices', Nature, vol. 264, no. 1, pp. 746-748.

[5] Bach-y-Rita, P 2003, 'Seeing with the brain', International Journal of Human Computer Interaction, vol. 15, no. 2, p. 285–295.

[6] Laurent, R, Olivier, C, Colline, P,Dai, T, Annick, V, Anne, B, Claude, V &Anne, VG 2005, 'Cross-modal activation of visual cortex during depth perception using auditory substitution of vision', NeuroImage, vol. 26, no. 1, p. 573– 580.

[7] Amedi, A,Stern, W, Camprodon, J & Bermpohl, F 2007, 'Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex', Nature Neuroscience, vol. 10, no. 1, p. 687–689.

[8] Bach-y-Rita, P 2004, 'Tactile sensory substitution studies, Annals of New York Academic Sciences, vol. 1013, no. 1, p. 83–91.

[9] Hadjikhani1, N & Per, R 1998, 'Cross-Modal Transfer of Information between the Tactile and the Visual Representations in the Human Brain: A Positron Emission Tomographic Study', The Journal of Neuroscience, vol. 18, no. 3, p. 1072–1084.

[10] Humphrey, N 1992, A History of the Mind: Evolution and the Birth of Consciousness, New York: Simon and Schuster.

[11] O'Regan, JK 2011, Why Red Doesn't Sound Like a Bell: Understanding the feel of consciousness, Boston: Oxford University Press.

[12] Vincent, W 1998, Perceptual Constancy: Why Things Look as They Do, Cambridge: Cambridge University Press.

[13]Koch, C 2004, The Quest for Consciousness: A Neurobiological Approach, Englewood: Robert & Company plublishers.

[14] Way, T & Barner, K 1997, 'Automatic visual to tactile translation, part I: human factors', IEEE Transactions on Rehabilitation Engineering, vol. 1, no. 5, pp. 81-94.

[15] Haigh, A, Brown, DJ Meijer, P & Proulx, MJ 2013, 'How well do you see what you hear? : The acuity of visual-to-auditory sensory substitution', Frontiers in psychology, vol. 4, article. 330.

[16] Tononi, G 2012, PHI: A voyage from the brain to the soul, New York: Pantheon Books.

[2] http://www.cs.berkeley.edu/~lbourdev/poselets/

[17]Berkeley, G 1709 (2008), An Essay towards a New Theory of Vision, New York: Aaron Rhames.
[18] Ward, J & Wright, T 2012, 'Sensory substitution as an artificially acquired synaesthesia', Neuroscience and Biobehavioral Reviews, vol. 1, no. 1, p. 11–17.
[19] Amedi, A, *Amir Amedi's Lab*, *The Hebrew University of Jerusalem*. Available: http://brain.huji.ac.il/site/em.html. [Accessed 12 March 2013].
[20] Ptito, M & Desgent, S 2004, Sensory input-based adaptation and brain architecture, Cambridge: Cambridge University Press.
[21] Maurice, P Solvej, M,Albert, G & Ron, K 2005, 'Cross-modal plasticity revealed by electro-tactile stimulation of the tongue in the congenitally blind', Brain, vol. 128, no. 3, p. 606–614.
[22] Bach-y Rita, P, Collins, C, Saunders, F & White, B 1969, 'Vision substitution by tactile image projection', Nature, vol. 221, no. 1, p. 963–964.
[23] Visell, Y 2009, 'Tactile sensory substitution: Models for enaction in HCI', Interacting with Computers, vol. 21, no. 1, p. 38–53.
[24] Raymond, M 1975, 'An Audio Display for the Blind', IEEE Transactions on Biomedical Engineering, vol. 23, no. 2, pp. 144-154 .
[25] T. Bower, "Blind babies see with the ears," newScientist, pp. 255-260, 3 February 1977.
[26] Kay, L 1996, 'Ultrasonic spectacles for the blind, in Conference of Sensory Devices for the Blind, London, UK.
[27] Heyes, D 1983, 'Human navigation by sound', Physics in Technology, vol. 14, no. 2, pp. 68-75.
[28] Meijer , P 1992, 'An Experimental System for Auditory Image', IEEE Transactions on Biomedical Engineering, vol. 39, no. 2, pp. 112-121.
[29] Capelle, C & Trullemans, C 1998, 'A Real-Time Experimental Prototype for Enhancement of Vision Rehabilitation Using Auditory Substitution', IEEE Transactions on Biomedical Engineering, vol. 45, no. 10, pp. 1279-1293.
[30] Capalbo, Z & Glenney, B 2009, 'Hearing color: radical plurastic realism and SSDs', in the 5th Asia-Pacific Computing and Philosophy Conference, Tokyo, Japan.
[31] Winlock, T, Christiansen, E & Belongie, S 2009, 'Toward Real-Time Grocery Detection for the Visually Impaired, in Workshop on Computer Vision Applications for the Visually Impaired, San Francisco, US.
[32] Bujacz, M 2010 , 'Representing 3D scenes through spatial audio in an electronic travel aid for the blind' Technical University of Lodz, Faculty of Electrical, Electronic, Computer and Control Engineering, Łódź.
[33] Ribeiro, F 2012, 'Auditory augmented reality: Object sonification for the visually impaired', in IEEE 14th International Workshop on Multimedia Signal Processing (MMSP), Banff, Canada.
[34] Andrea Bochelli Foundation, *MIT Fifth Sense Project: Providing the Functions of Sight to Blind People* . Available: http://people.csail.mit.edu/seth/misc/bocelli.html. [Accessed 1 March 2013].
[35] Bologna, G, Deville, B, Gomez, J & Pun, T 2010, 'Toward local and global perception modules for vision substitution', Neurocomputing, vol. 74, no. 8, p. 1182–1190.
[36] Gomez, J, Bologna, G & Pun, T 2011, 'Multisource sonification for visual substitution in an auditory memory game: one, or two fingers' in The International Conference on Auditory Display ICAD, Budapest, Hungary.
[37] Davis, C, *The CIPIC Interface Laboratory Home Page*. Available: http://interface.cipic.ucdavis.edu/. [Accessed 1 February 2012].
[38] Begault, R 1994, 3-D Sound for Virtual Reality and Multimedia, Boston: AP Professional.
[39] Brown, C & Duda, R 1998, 'A structural model for binaural sound synthesis', IEEE Transactions on Speech and Audio Processing, vol. 6, no. 1, pp. 100-105.

[40] Gomez, J, Bologna, G, & Pun, T 2012, 'Non-visual-cueing-based Sensing and Understanding of Nearby Entities in Aided Navigation', in The 14th International ACM Conference on Computers and Accessibility ASSETS, Boulder, US.
[41] Gomez, J, Bologna, G & Pun, T 2012, 'Spatial Awareness and Intelligibility for the Blind: Audio-touch Interfaces', in The ACM SIGCHI Conference on Human Factors in Computing Systems CHI, Austin, US.
[42] Javed, O, Ali and, S & Shah, M 2005, 'Online detection and classification of moving objects using progressively improving detectors', in Computer vision and pattern recognition conference CVPR, San Diego, US.
[43] Shiuh-Ku, W, Chung-Ming, K & Shu-Kang, T 2006, 'Video object tracking using adaptive Kalman filter, Journal of Visual Communication and Image Representation, vol. 17, no. 6, p. 1190–1208.
[44]Kalal, Z, Matas, J & Mikolajczyk, K 2009, 'Online learning of robust object detectors during unstable tracking', in On-line Learning for Computer Vision Workshop, Kyoto, Japan.
[45] Kalman, RE 1960, 'A new approach to linear filtering and prediction problems', Journal of Basic Engineering, vol. 82, no. 1, p. 35–45.
[46] Guangcheng, Z, Xiangsheng, H, Stan, Z & Li, Y 2005, 'Boosting Local Binary Pattern (LBP)-Based Face Recognition' Advances in Biometric Person Authentication, vol. 3338, no. 1, pp. 179-186.
[47] Ward, J &Graham, M 2011, 'Evolving the ideal visual-to-auditory sensory substitution device using interactive genetic algorithms', Journal of human, animal, and machine perception, vol. 2, no. 8, p. 37–41.
[48] Auvray, M,Hanneton, S & O'Regan, K 2007, 'Learning to perceive with a visuo-auditory substitution system: localisation and object recognition with 'The vOICe'', Perception-London, vol. 36, no. 3, pp. 100-108.
[49] Gaver, W 1986, 'Auditory icons: Using sound in computer interfaces', Human-Computer Interaction, vol. 2, no. 1, pp. 167-177.
[50] Poiriera, C, De Voldera, A & Scheiberb, C 2007, 'What neuroimaging tells us about sensory substitution', SinceDirect, vol. 2, no 37, p. 1064–1070, 2007.