A computer-vision based sensory substitution device for the visually impaired (See ColOr)

Gomez Valencia, Juan Diego

UNIVERSITÉ DE GENÈVE
Département d'Informatique

FACULTÉ DES SCIENCES
Professeur Thierry Pun
Dr. Guido Bologna

A computer-vision based sensory substitution device for the visually impaired (See ColOr)

THÈSE

présenté à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention informatique

par
Juan Diego GOMEZ VALENCIA
de
Neira (COLOMBIA)

Thèse N° 4642

GENÈVE
Repro-Mail - Université de Genève
2014

UNIVERSITÉ
DE GENÈVE
FACULTÉ DES SCIENCES

Doctorat ès sciences
Mention informatique

Thèse de *Monsieur Juan Diego GOMEZ VALENCIA*

intitulée :

"A Computer-vision Based Sensory Substitution Device for the Visually Impaired (See Color)"

La Faculté des sciences, sur le préavis de Messieurs T. PUN, professeur ordinaire et directeur de thèse (Département d'informatique), G. BOLOGNA, docteur et codirecteur de thèse (Département d'informatique), S. MARCHAND-MAILLET, professeur associé (Département d'informatique), Ch. JOUFFRAIS, docteur (Laboratoire commun Institut de Recherche en Informatique de Toulouse/ Institut des Jeunes Aveugles - Lab on Assitive Tech for the Visually Impaired, Université Paul Sabatier, Toulouse, France) et Madame E. PISSALOUX, professeure (Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie et Centre National de la Recherche Scientifique, Paris, France), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 11 février 2014

Thèse - 4642 -

Le Doyen, Jean-Marc TRISCONE

N.B.- La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Acknowledgments

The dreams of PhD begin with a problem, an equation, perhaps an algorithm, or, back in the alma mater, with a girlfriend to whom I also offer this thesis: Johanna Muñoz. Likewise, Luis Alfonso Muñoz not only served me as an inspiration but, helped me walk the very first step which this trek began with. To both of them I owe my humblest thankfulness: the mother and the grandpa of my daughter. Let me thank as well Thierry Pun for giving me the opportunity to get this thesis done by his side. Certainly, his fine sense of humor made things easier for me. Beyond the admirable scholar, Thierry is a good soul, neat and honest in his actions. I will thank endlessly Guido Bologna for priceless advising and help, but overall, for those fruitful and enriching chats spanning from Wall Street to ineffable ultraviolet colors only perceivable by the bee eye.

I thank Diego for a series of 1327 e-mails that we exchanged since 2010 when I started my PhD. Thanks buddy for not leaving me alone within the loneliness. As our song goes: *tantos siglos, tantos mundos, tanto espacio... y coincidir!* Thanks to my mother Consuelo for the everlasting candle in her kitchen's corner. Thanks mom for the light. All my gratitude goes also to three key people: Hugo Gomez for offering me a cup of ayahuasca brew every year, the scared vine from which I've learned a great deal. Thanks 'taita' Hugo for the remedy. To Cata Saffon for being my best friend, the highest personal achievement of this thesis. Thanks Cata for teaching me how life may change for the better. I love you both, you and your daughter. And Juliana for her love and modesty, for those splendid nights, reflecting via skype on Buddhism and consciousness. We both are walking down the middle path towards compassion and wisdom.

Oddly enough, I do have quite a bit to thank to Julio Sanchez Cristo. Even though I've never met him personally, we shared all these years of phding through a radio receiver. Julio kept my heart attached to my home country by means of a microphone, a poem, a song, news. Reason why this acknowledgment section began by copying his speech "the dreams of radio", same he used to collect his radio award Rey de España.

Last but not least, I cannot thank enough my lovely daughter for being the fundamental engine. Like Charles Bukowski, many were the mornings that I woke up and thought: *I'm just not gonna make it*. Yet, in my haste to find her a better future and set my best example, I was determined to reach the end, no matter what I had to lose in the go. Lacking Sofia, none of this would have made sense. Luckily enough, it wasn't so. Although, during these, her earliest years, she sadly had to miss me. To my beloved one, my all, I essentially dedicate this thesis as a token of love:

I love you my babe and I always will.

Agradecimientos

Los sueños de doctorado comienzan con un problema, tal vez una ecuación, un algoritmo o con una novia cómplice –allá en el alma mater– a quien también dedico esta tesis: Johanna Muñoz. En la misma línea, Luis Alfonso Muñoz no solo me sirvió como inspiración, si no que me ayudó a dar ese primer paso con que comenzó el viaje. A ellos mí más humilde Gracias: la madre y el abuelo de mi hija. A Thierry Pun quiero agradecer la oportunidad que me brindó de realizar está tesis a su lado, en su laboratorio. Brillante académico Thierry además es un gran hombre, pulcro y honesto en su actuar. A Guido Bologna, le voy a gradecer siempre su tutoría y su ayuda, pero sobre todo, esas enriquecedoras charlas que podían variar desde Wall Street, hasta como imaginar inefables colores ultravioletas solo perceptibles con el sistema visual de una abeja.

Le agradezco a Diego por una serie de 1327 correos electrónicos que hemos compartido desde 2010 que comencé mi doctorado. Gracias por no dejarme nunca solo en esta soledad. Cómo dice la canción que le regalé: tantos siglos, tantos mundos, tanto espacio... y coincidir! A mi madre Consuelo le agradezco por esa veladora encendida en un rincón de su cocina. Gracias por la luz mamá. Voy a agradecerle a tres personas muy especiales. A mi tío Hugo Gómez por brindarme cada año ese trago de yagé, el vejuco del alma, del que tanto he aprendido. Gracias por el remedio taita. A Cata Saffon por convertirse en mi gran amiga, el logro personal más lindo que me deja esta tesis. Gracias Cata por la lección de cambio, te quiero a ti y a tu hija. A Juliana por la compañía y por esas refinadas noches de reflexión y filosofía, vía Skype, sobre Budismo y mente. Gracias Juli por tu amor y tu sencillez, los dos vamos avanzando juntos por el camino medio, hacia la compasión y el despertar de la conciencia. Gracias al budismo por enseñarme tanto de la vida, y viceversa.

Este es un muy particular agradecimiento, a un hombre que admiro, a Julio Sánchez Cristo, a quien no conozco, pero con quien he compartido todos estos años a través de un radio. Julio me ha mantenido el corazón atado a mi país por medio de un micrófono, un poema, una canción, una noticia. Es por eso que estos agradecimientos los comencé con el encabezado de su discurso “los sueños de radio”. Aquel que usó para recibir su premio de periodismo Rey de España.

A mi hija Sofía, no puedo agradecerle con palabras ser el motor de todo esto. Igual que Charles Bukowski, muchas mañanas abrí los ojos y pensé: ‘no voy a lograrlo’. Pero para asegurarle un mejor futuro y sobre todo darle un mejor ejemplo. Mi determinación fue llegar a la meta sin importar lo que perdiera en el camino. Si Sofía me hubiese faltado, nada de esto habría tenido sentido. Pero no fue así, no me faltó, aunque en estos, sus primeros años, yo le tuve que faltar a ella. A mi niña en primer lugar le dedico esta tesis.

Te amo mi bebé, gracias.

Abstract

Visually disabled individuals face significant difficulties in interacting comfortably with their environment as their mobility and cognition of the world are dramatically diminished. The main aim of the study carried out and reported in this thesis is to build a prototype for visual substitution in the interest of helping the blind and the visually impaired. Overall, we introduce *See ColOr*, whose name stands for seeing colors with an orchestra, as a Sensory Substitution Device (SSD) pursuing a long-standing goal in visual rehabilitation, namely *'seeing through the ears'*. In principle, quite like related works, we adhere to neuroplasticity theories, which broadly imply that visual consciousness can be elicited bypassing the eyes through the auditory sensory pathway. Accordingly, this manuscript begins with a review of the physiology of vision and the literature related to SSD, to shed light on the development of this scientific endeavor through the last five decades.

See ColOr is a SSD that proposes a unique code to map optical colors into instruments sounds, allowing visual stimuli to be conveyed as audio cues. It uses a head-mounted 3D-camera, a tactile interface (tablet) and bone-phones to transmit sound via bone stimulation without blocking out the ears. In addition, unlike typical approaches, *See ColOr* uses also Artificial Vision to simulate higher-level or cognitive aspects of vision: object recognition, face detection, text reading, and context awareness to prevent users from bumping into unexpected obstacles. Accordingly, this thesis contributes on a variety of research topics, such as: simultaneous sonification of color and depth using spatialized virtual sound sources; efficient processing of range images; and also relevant aspects of human computer interaction and haptic-audio trajectory playback. At the experimental level, we contribute with systematical evaluations that involve both, mobility and orientation, which is a major lack in the state of the art. By and large, *See ColOr* proved to allow its users grasping visual information of the world out of which they can derive: spatial awareness, ability to find someone, location of daily objects, and skill to walk safely avoiding obstacles. In this way, we largely answer several research questions about how reliably humans can perceive color through sound, and how much visual information is actually codable into audio. As a matter of fact, we can "input a scene" in someone's mind through the ears, in a few minutes with a precision of centimeters.

With regard to the inclusion of computer vision into our approach, we also review plausible ideas on the ontological nature of vision, to argue that visual perception is unlikely attainable by means of today SSDs. The hypothesis central to our approach is that by combining sensing, computation and interaction, as our final experiments confirmed, we are bound to achieve a wearable device more functional, learnable and practical, capable of producing reliable knowledge about the physical world. The outcome of this thesis is then a smart *See ColOr* that can be adapted to multiple tasks. Moreover, as a last novelty, we get rid of the tablet with the concept of *tactile augmented reality*, which allows users interacting with the system only through hand gestures. This prototype, affordable as it is compared to retinal

implants, was tested in a developing country where the access to this kind of technology is rather little. This is a venue where, in nearly 90% of the cases, blindness prevents people from working and drops their life expectancy down to 1/3. The enthusiasm and blissfulness of this South American community when we reached them with *See ColOr*, will linger in our memories through a series of videos recorded to conclude this thesis.

to Sofia...

"Knowledge can be communicated,
but not wisdom. One can find
it, live it, do wonders through it,
but one cannot possibly
communicate or teach it."

Siddhartha, Herman Hesse (1922)

Table of Contents

Acronyms and Abbreviations

2bitBP: 2 bit binary pattern.

AI: Artificial intelligence.

ANN: Artificial neural network.

BoW: Bag of words.

CPU: Compute processing unit.

CVML: Computer vision and multimedia laboratory.

CVS: Computer vision system.

Earcon: a brief, distinctive sound representing a specific event or convey other information.

EEG: Electro encephalogram.

FFT: Fast Fourier transform.

fMRI: Functional magnetic resonance image.

FOV: Field of view.

FPGA: Field-programmable gate array.

FT: Fourier transform.

HCI: Human computer interaction.

HRIR: Head-related impulse response.

HRTF: Head-related transfer function.

HSL: Hue, saturation, lightness.

HSV: Hue, saturation, value.

IID: Inter-aural intensity difference.

IR: Infra-red.

ITD: Inter-aural time delay.

KF: Kalman filter.

NDT: Normal distribution transform.

OCR: Optical character recognizer.

PCA: Principal components analysis.

PET: Positron emission tomography

RGB: Red, green, blue.

SDK: Software development kit.

See ColOr: Seeing colors with an orchestra.

SIFT: Scale-invariant feature transform.

Soundscape: a combination of sounds that forms or arises from an immersive environment.

Spearcon: Speech-based earcons.

SS: Sensory substitution.

SSD: Sensory substitution device.

SURF: Speeded up robust features.

SVM: Support vector machine.

TDU: Tongue display unit.

ToF: Time of flight.

TUIO: Table-Top user interfaces objects.

TVSS: Tactile visual sensory substitution.

USB: Universal serial bus.

WHO: World health organization.

1 INTRODUCTION

The World Health Organization estimates the world blind population at 39 million persons, which roughly corresponds to 0.56% of the total world population. More precisely this represents an incidence ranging from 0.5% to 1.4% in the developing countries and of 0.3% in the whole of the industrialized countries [**1**]. As for low-income countries, in nearly 90% of the cases a blind individual can no longer work and his/her life expectancy drops down to 1/3 that of a matched peer, in age and health. Back to the global picture, blindness is likely to double in the next 15 years because of ageing. In the world, people aged over 60 account for 58% of blind. In Switzerland for instance, 10000 people are affected by blindness and 80% of them are more than 60 years old. Fortunately, on a worldwide scale approximately 50% of blindness could be prevented. Nonetheless, without effective and major intervention, the projected increase in global blindness to 76 million by 2020 will be regrettably reached. [**1**]

At large, blind individuals are adept at traveling with help of traditional mobility aids (i.e. white canes and guide dogs). The limitations of these tools, however, become apparent in numerous daily life situations, creating a strong urge to seek aid from others. Accordingly, the exploration of new environments turns out particularly demanding. Also, when looking for unfamiliar destinations, it is very challenging for them to handle unexpected needs or notice serendipitous discoveries that might arise on their way. Despite the use of traditional assistance, blind individuals still miss a great deal of information of the environment that sighted people may take for granted. Last but not least, their other perceptual capacities may be further lessened by the focus needed for mobility and orientation tasks poorly assisted. This eventually lowers their sense of independence and dignity.

Nowadays, state-of-the-art retinal implants are intended to restore some functional vision lost after damage of the photoreceptors, the most common cause of blindness (e.g. retinitis pigmentosa and macular degeneration). These implants benefit from the fact that both optical nerve and visual cortex remain undamaged, so that by electrically stimulating fibers in the optical nerve, wasted photoreceptors may be bypassed. Roughly, a small camera captures a video that is coded and sent wirelessly to an implanted electrode array, emitting pulses of micro-electricity towards the brain. Clinical trials unfortunately reveal that these neuroprothesis still suffer from very limited resolution (i.e. 6×10 electrodes). Indeed, implanted patients reported having perception of mere light patterns devoid of legibility and needing complex interpretation. Therefore, basic visual tasks remain challenging or impossible for them, such as objects identification, navigation in unknown environments or detection of surrounding objects or persons identity [**2**] [**3**]. This added to clinical risks of invasive treatments (let alone high prices), has augmented the skepticism of many that await affordable solutions, less risky, and more efficient.

This scenario has given rise to a proliferation of context-aware research and development (e.g. Electronic Travel Aids, Global Positioning Systems and Geographical Information Systems, Sensory Substitution Devices 'SSDs'). Particularly, SSDs are made up of an optical sensor coupling a processing device that systematically converts visual features into tactile or auditory responses [**4**]. Thus, the goal here is convey visual information to the sense of hearing (or touching). The central idea of SSDs is rooted in the concepts of multisensory perception and cross-modal transfer[1] [**5**], holding that perception entails interactions between two or more different sensory modalities. This implies that areas of the brain typically associated to the processing of inputs from a specific sensory pathway, may be activated by other senses after robust training. In principle, the advantages of this sort of devices are clear: noninvasive technology at relatively low cost.

While generally promising, there are still a number of significant gaps in our understanding of the HCI issues associated with SSDs. Overall, a subject that remains relatively uncertain relates to the usability in real scenarios. The underlying problem is that the capacity of information transfer of the human eye reaches 1000 Kbps [**6**]. Whereas, senses intended as substitutes can hardly reach 10 Kbps at most (i.e. hearing) [**6**]. Thus, even though a cross-modal transfer may apply, it is hard for mapping systems to overcome the large sensory mismatch between visual perception and other sensory pathways. Accordingly, many SSDs very often suffer from either loss of great deal of visual information, or illegible representation thereof. In practice, this fact dramatically diminishes their usability.

In this view we put forward See ColOr, a mobility assistance device for the blind aimed at making a step further toward their independent mobility. See ColOr is a non-invasive mobility aid that uses the auditory pathway to represent a RGB-D (red, green, blue and depth) stream in real-time. In principle, See ColOr encodes points of captured pictures into spatialized musical instrument sounds, so as to represent color and location of entities. More specifically, these points are represented as directional sound sources, with each emitted instrument depending on color. Also, the depth is represented by the rhythm of the sound. The strategy for selecting points may be either automatic selection of the center of the picture, or customized selection by tapping on a tactile tablet within which the picture is presented. Ultimately though, See ColOr attempts at providing a hardware-free interaction, thus the user will only need to point with the fingers spots in the real space in order to sonify them (see Haptic-based Interfacing and Discussion).

Since the aforementioned functionalities are limited to describe local portions of an image using low-level features such as color and depth, they foster micro-navigation[2] [**7**] of entities by allowing selective exploration, discovery of points of interest, comparisons, and, in general, to enjoy a greater sense of independence. Nevertheless, they might fail to reveal cognitive

[1] is perception that involves interactions between two or more different sensory modalities.
[2] micro-navigation is concerned with detecting and avoiding obstacles while walking through immediate environment.

aspects which often determine regions of interest within a picture. Accordingly, See ColOr also uses computer vision techniques to process higher visual features of the images in order to produce acoustic virtual objects [**8**]. Actually, we recognize and then sonify objects that do not intrinsically produce any sound, with the purpose of revealing their nature and location to the user. Overall, this allows the blind noticing serendipitous discoveries; seeking a specific target; and avoiding obstacles.

1.1 Our thesis

Over the last four decades a proliferation of SDDs has emerged out of the interest among the research community in aiding the blind. In particular, the main challenge faced by auditory-based systems is the overcoming of the information bandwidth mismatch between the complex visual spatial input and the auditory output (i.e. sensory overload). Therefore, research on this topic has been focused on the transduction of low-and-middle level visual features into the audio cues, such as brightness, contrast, color, spatial awareness, depth etc. Arguably these approaches have failed to build a model replicating higher levels of the visual system. Hence, they completely neglect cognitive aspects which often determine regions of interest in the visual field or information subject to top-down knowledge [**9**].

Vision is a phenomenon that entails both, sensation and perception [**10**], [**11**], [**12**]. 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 out of 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 [**6**]. Yet, senses intended as substitutes can hardly reach 10^4 bps at most (i.e. hearing) [**6**]. 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 or enhanced, in order for the substitution to be: practical, fast, accurate, and let users act as though they were actually seeing [**14**]. Visual perception is a holistic phenomenon that emerges from complex information unlikely to be encoded into hearing (shapes, perspectives, color, position, dis-

tance, texture, luminance, concepts etc.) [**10**], [**11**], [**15**], [**16**], [**17**]. In fact, 'normal' vision is itself constrained by top-down knowledge that produces the kind of information that sighted individuals achieve typically without conscious effort [**12**], [**13**], [**14**]. Our **Thesis** 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 extend such an approach to the use of computer vision and image processing to deal with high-level information that usually surpass the bandwidth of audio. Whether this strategy will lead us to an SSD more learnable, practical, easy to interact with, and chiefly functional? Is a fundamental research question underlying this work.

In short, we do believe that the coding into sound of basic visual cues (e.g. color and depth) accompanied by computational methods that model higher perceptual levels of the visual system will lead us to a SSD: functional, ease to use, and suitable for mobility and exploration tasks "See ColOr". This will be done by leveraging innate human capacity for sound distinction (and localization) as well as the usefulness of computer vision methods to synthesize human sight ability.

1.2 Research questions, scope of this thesis, and main contributions

In stating our thesis several research question arise, such as:

- Can humans reliably perceive color and depth by senses other than sight?
- Are cutting-edge technology and state-of-the-art methods capable of accurately representing a visual scene in someone's mind, through sound?
- Can we, nowadays, engineer a system that allows the blind to behave nearly as the sighted individuals do? In which extent?
- Does it exist an ideal way to represent complex visual elements through sound?
- Is there any method leading to optimal interaction of blind users and aiding systems based on touch screens?

Throughout this document the development of studies, surveys, experiments, implementations, statistics and comparisons, will provide insight into answering the aforementioned questions in order to support our thesis. As a matter of fact, the conclusions of this document will relate our work to the answers we pursuit. Generally speaking, this work aims at providing a mobility aid prototype with fundamental research basis. However, due to technical limitations we will constrain this approach to indoor environments. Also, we want to note that a fully developed system ready for commercialization ends is not the target of this work. Rather, we offer scientific guidelines and meaningful testing for further adaptations to end-user systems. Essentially, we will follow four lines of investigation: sonification, range imaging, haptic interfacing and computer vision. Our contribution on sonification is the developing of a sonic code that maps colors and depth into musical instruments sounds using spatialized

audio. Besides, experimental grounds on simplified scene sonification will be provided in arguing against usual methods such as soundscapes[3].

We will also contribute in this thesis with a framework for the coupling optical sensors in the context of range and color image registration. Computer vision methods for object recognition will be implemented and tested in this work. Importantly though, we do not attempt at improving the state of the arte in the field of artificial vision, therefore we will make use of available methods. Also, a tactile interface will be implemented followed by robust research and testing to assess how this could lead blind users to achieve better insight into the visual world. All these topics will be condensed into our prototype robustly tested through series of experiments conducted with blindfolded and blind individuals. It is worth noticing that although SDDs is a broad field of research, the state-of-the-art review given in this document will be limited to those technologies based on auditory substitution of vision. In short, the main contributions of this thesis can be regarded as follow:

- A functional prototype for aiding the visually impaired in exploration and mobility.
- A sensorial coding of color and depth into sound.
- An alerting method simple, yet fairly efficient to maintain the blind user's safety.
- An optimal haptic interface to mediate information between blind users and sound-represented visual environments.
- Meaningful insight into sonification methods of visual cues and worthwhile ideas to correlate sound and vision into the brain.
- An implemented framework (Matlab-based) to integrate low-level and high-level visual features into SSDs.
- Implementation of state-of-the-art computer-vision-based techniques oriented to blind assistance.
- A Braille-like (based on tactile exploration) text recognition method based on deep learning.
- The concept and implementation of Tactile Augmented Reality.
- An efficient method for registration of depth and color sensors to enhance performance in aiding the blind.
- A method for orthographic camera simulation (ortho-kinect) oriented to scene simplification in audio representation.
- Robust experimental basis involving end-users both, blindfolded and blind people.
- Meaningful discussion on the problem of visual substitution aids, covering historic aspects, current issues, future challenges and novel ideas.
- A community-based research approach to reach out developing-country population.

These contributions have been reflected in a number of publications issued during the thesis work and cited in the bibliography , namely: [**18**], [**19**], [**20**], [**21**], [**22**], [**23**], [**24**], [**25**],

[3] is a sound or combination of sounds that forms or arises from an immersive environment, e.g., the natural sound of a jungle (birds, wind, cricks etc.).

[**26**], [**27**], [**28**], [**29**], [**30**]. Also, we added an independent list with these publications in Appendix A. In addition, out of this thesis four bachelor projects emerged under advice of the author. This thesis also contributed with meaningful help to one master project:

- *Kinect-based Autonomous mini Robot Car* by **Mikhail Chantillon**. Bachelor in computer science, University of Geneva (2011).
- *Kinect-based object detection for visually impaired people* by **Sinan Mohamed**. Bachelor in computer science, University of Geneva (2011).
- *Kinect-based text detection and recognition in the interest of accessibility* by **Thomas Dewaele**. Bachelor in computer science, University of Geneva (2012).
- *Computer Vision for Mobility Aids: On the detection of ground surface changes* by **Bruno Barbieri**. Bachelor in computer science, University of Geneva (2012).
- *Follow me: A computer-vision-based programing* of a NAO robot by **Sheu Wen-Ru**. Master in computer science, University of Geneva (2013).

1.3 Thesis structure

Chapter 2 (BACKGROUND) of this document offers the neurological basis of vision that will support both, our SSD See ColOr and our idea of adding computer vision to classical SSDs. Following, a critical extensive review of the state of the art in SSDs is also provided in this chapter. Chapter 3 (SEEING COLORS WITH AN ORCHESTRA), in turn, fully describes the implementation of See ColOr across four lines of research, namely: sonification (3.3 Sonification), image registration (3.4 Efficient registration of range and color images), haptic interfacing (3.5 Haptic-based Interfacing) and computer vision (Computer-vision-based visual substitution). Subsequently, in chapter 4 (EXPERIMENTS), we validate See ColOr by means of a series of experiments conducted with both, blindfolded sighted and blind individuals. Also, chapter 4 closes with relevant arguments to justify the functionality of See ColOr. The final chapter of this document (chapter 5, DISCUSSION AND CONCLUSIONS) concludes this document with a general summary, addressing as well research questions, future work, user's feedback and the lessons we learned from our own work. Importantly, while this thesis provides a full chapter dedicated to experiments (chapter 4), all the research depicted in chapter 3 accounts for experimental basis too, as shown in Table 1-1.

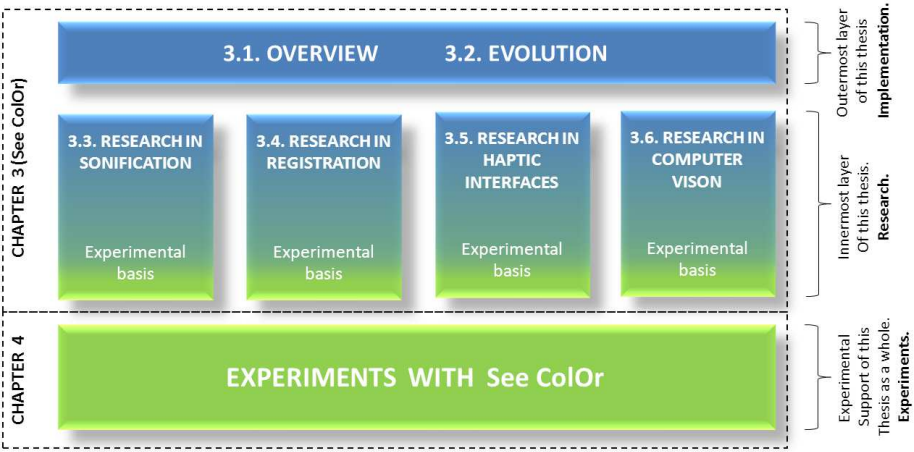


Table 1-1. Structure of chapters 3 and 4 of this document.

2 BACKGROUND

This chapter is intended to track vision down to its lair in the brain, giving insights into its most intricate ontological aspects. We study both, the physiology and the psychology of vision in humans with two aims (2.1.1 Vision and Blindness). Firstly, to review the neurological grounds of sensory substitution that ultimately support our SSD (2.1.2 Multisensory perception, 2.1.2 Multisensory perception and 2.1.3 Cross-Modal Transfer and Brain Plasticity). Secondly, to make it clear that visual consciousness takes more than visual sensations encoded into sound by todays SSDs. This latter is a fundamental observation on which our thesis of adding Computer Vision to classic SSDs entirely relies. An extensive critical review of the state of the art in SSDs concludes this chapter, so as to provide the reader with an account of what science has done so far to achieve the goal: seeing without the eyes (2.3 Sensory Substitution Devices SDDs (State of the art)).

2.1 About the brain

2.1.1 Vision and Blindness

Visual perception

Visual perception is the ability to interpret the surrounding environment by processing information that is contained in visible light. By and large, there are four main components that make it possible for human to have a subjective, conscious visual experience (i.e. qualia, Bach-y-Rita *et al.* [**31**]), namely: light, eyes, optical nerve, and the brain. The whole phenomenon of vision can be roughly summarized as follows: the light reflected by objects in the world reaches the eyes through the cornea (the outermost layer of the frontal eye). This light then strikes the retina which is the inside surface of the back of the eye ball. When light strikes the retina, it triggers activity (impulses) in photoreceptors that synapse with the axons of the optical nerve. Nerve impulses are then sent to the brain (through the optic pathway or nerve) to be interpreted as visual images. Thus, about 2 billion neurons in the visual cortex located in the back of the brain (within the occipital lobe) start firing through 500 trillion synapses. From this point onward, the visual experience is no longer traceable by today's scientific methods. We just know that a visual image is then produced and projected back into space to clothe the observed object, so that this object begins to exist in the visual consciousness of the person [**11**]. Here follows an endless debate on consciousness embracing philosophy, religion and science (the mind-body problem [**11**]): "I" am the observer of this occurrence (visual experience) or an integral part of it?

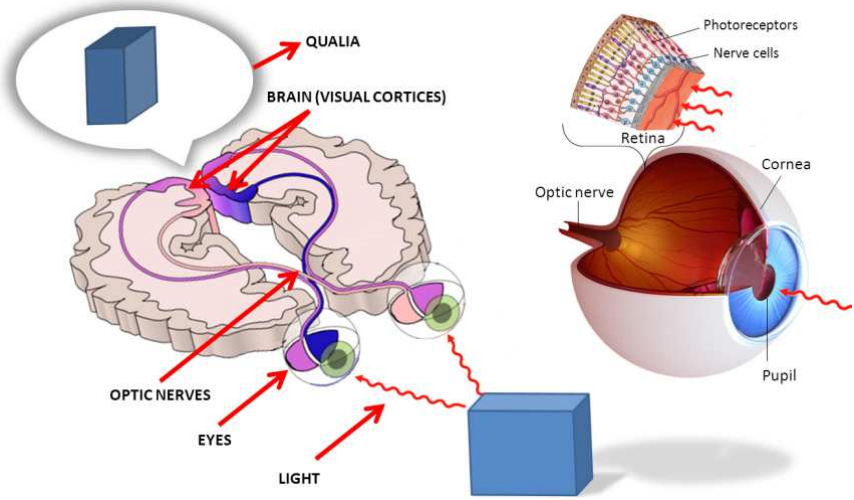


Figure 2-1. Emergence of a visual perception (qualia) through the visual system defined by its main parts (eyes, optic nerves and brain 'visual cortex'). The eye has many of the features of a camera lens to focus exterior light entering through the cornea. Right lights input are regulated by the pupil by either expanding or shrinking. After crossing the eye ball from side to side, light forms the visual image back in the retina. Importantly though, the image focused in the retina is inverted top-to-bottom and reversed right-to-left. Following, photoreceptor cells of the retina get stimulated and send those stimuli to the brain down the optic nerve. Finally, the visual cortex enters bioelectrical stimulation and composes the image in a coherent way so we can see. Note also that visual cortex activity is not only caused by optic nerve's stimuli. Memories, visual imaginary, thinking and dreaming among others are also elicitors of bioelectrical stimulation in this area. (Modified from wwww.2-sight.eu/)

Blindness

Blindness occurs when any of the elements involved into visual experiences is missing or defective. Note that even the lack of exterior light causes a temporal "blindness", because typically we cannot see in the darkness. More formally, blindness is defined as the lack of visual perception due to either physiological (the eye blind) or neurological factors (the mind blind):

The blind eye

According to WHO [**1**], the most common causes of blindness around the world are: cataracts (47.9%), glaucoma (12.3%), age-related macular degeneration (8.7%), corneal opacity (5.1%), diabetic retinopathy (4.8%), childhood blindness (3.9%), trachoma (3.6%), and oncho-

cerciasis (0.8%) [**1**]. These conditions may be acquired through illnesses, genetic disorders, injuries, poisoning (chemical toxins), aging, infections, etc. All of them, nonetheless, cause blindness due to damaging the eyes (or any of its functional parts such as cornea, retina, pupil etc.), and in some cases the optic nerve. Onchocerciasis also known as river blindness, for example, is a devastating parasitic disease caused by infection by a filarial worm termed Onchocerca volvulus [**32**]. This disease provokes long-term corneal inflammation (keratitis) that in turn, leads to thickening of the corneal stroma until blindness [**32**]. The parasite is transmitted to humans through the bite of a black fly of the genus Simulium. Most infections are reported in sub-Saharan Africa (approximately 270,000 cases), though also in Yemen, Central and South America, cases have been documented [**1**], [**32**].

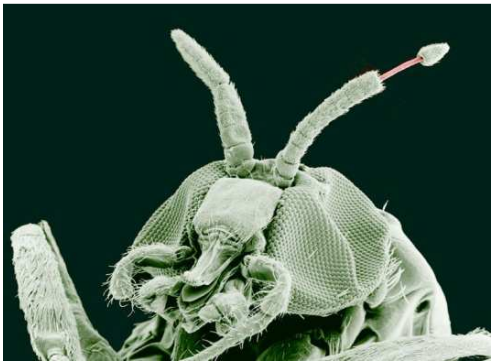


Figure 2-2. (taken from [32]) Adult black fly (Simulium yahense) with Onchocerca volvulus emerging from its antenna. Observed using conventional scanning electron microscopy. Magnified 100X.

It is a commonplace observation, however, that the brain (not the eyes) is in charge of the greatest deal of the visual experience. The eyes (and the optic nerves), while being important to vision, they have a role similar to that of a sensor (just acquiring information) rather than a processor (deriving sense out of information). In fact, we will see (later in this section, and in general in this thesis) that for eliciting a visual-like experience eyes and optical nerves could all be bypassed, but the brain. In this view, many scholars agree on saying that "we do not see with our eyes" [**33**], [**16**], [**15**], [**34**]. For instance, Johannes Kepler [**16**] was quoted as saying: "vision occurs through a picture painted on the dark surface of the retina. The eye is like a camera obscure, where the image is reversed. We don't see the world upside down nevertheless. This is because the eyes have little to do with the conscious visual experience; they are rather a door through which light enters the mind." Certainly, Kepler made a point in this statement. The retina, indeed, has rather low intervention in conscious vision: at its center, where the optical nerve takes its leave, it is not even sensitive to light, yet we do not see a hole in the middle of each sight [**16**]. And though the outer parts of the retina are blind

to color, we don't see a greyish vision field on its top. Further, the retina is constantly flickering due to imperceptible eye movements; nonetheless the conscious image remains stable.

In most of the cases (we will see exceptions here later), when individuals go blind, they do not actually lose the ability of seeing; rather they become incapable of convening external stimuli to the brain. Since the working of the brain is not affected, a person who lost the ability to retrieve data from their eyes could still create subjective images using: visual imagery, memories, and dreams among others. In this regard, O'Regan [**35**] says: "The only difference is that whereas imagining finds its information in memory, seeing finds it in the environment. One could say that vision is a form of imagining, augmented by the real world." As a remarkable example to this idea we can take Gian Paolo Lomazzo [**36**], an Italian painter, best remembered for his writings on art theory. He became blind at early age, yet he turned into a prominent art theoretician. His critics on art paintings were all based on oral descriptions, memories and tactile feedback from simple brushstrokes [**36**]. He also supervised the creation of master pieces that were first born (full of color and details) into his imagination. This was possible because his, was a kind of retinal blindness, one of the most common [**36**]. Metaphorically, his visual brain stayed locked in by his own eyes, so to speak.

Back to the present, at University of California Berkeley, Professor Jack Gallant provides meaningful insight into the brain to understand its workings regarding vision (**Figure 2-3**). Gallant *et al.* [**37**] have been able to reconstruct the visual experience of a subject out of his brain activity. They developed a new Bayesian decoder [**37**] that uses fMRI images from early and anterior visual areas to reconstruct complex natural images. In other words, they are building a dictionary that will enable accurate prediction of what the brain activity would be given a particular image, and vice versa [**37**]. Here they are getting to the last step where vision can be tracked in physical terms. The uncertainty comes when the visual cortex activity has taken place in the brain, since right in there lies the indistinguishable border between the physical workings of our nervous system and ultimate immaterial nature of vision. Yet, we can see clearly now that being the last step before reaching visual consciousness from external light, the brain plays an decisive role in this conversion. If someone can tell us at the end, what vision really is and what it is made of, that has to be the brain. As matter of fact, we will see in the sections to come that out of this path (from light hitting the eye to vision) many steps may be broken or removed but the last (the brain workings). The brain is more intimately related to vision than any other element in the visual system, including eyes. Actually, works such as that of Gallant tells us that while light is acquired through the eyes, it becomes vision just upon arrival to the brain.

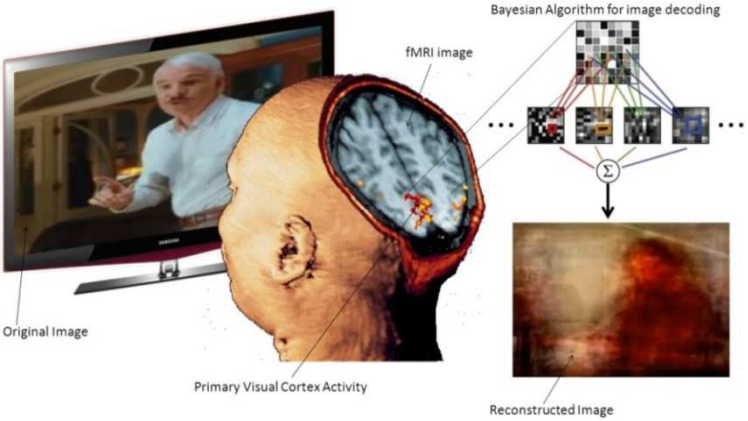


Figure 2-3. Image reconstruction based on brain activity. Images are translated into activity patterns of the primary visual cortex. So far, it is being exanimated just the part of the primary visual cortex that responds to little local visual features in the images such as, small edges, colors, short motions and textures. This part of the brain however does not account for a number of things such as what the objects are in the image. Consequently, the reconstructed image still differs much from the original one. Further progresses are expected soon. Note also that visual cortex activity is not only caused by worldly images. Memories, visual imaginary, thinking and dreaming among others are also elicitors of bioelectrical stimulation in this area.

Nowadays, there is more and more practical evidence of these brain-centered approaches to vision. Retinal implants, for instance, are intended to restore some functional vision lost after damage of the photoreceptors, another common cause of blindness (e.g. retinitis pigmentosa and macular degeneration [**1**]). These implants benefit precisely from the fact that both optical nerve and visual cortex, remain undamaged, so that by electrically stimulating fibers in the optical nerve, wasted photoreceptors may be bypassed. Roughly, a small camera captures a video that is codified and sent wirelessly to an implanted electrode array, emitting pulses of micro-electricity towards the brain. Clinical trials unfortunately reveal that these neuroprothesis still suffer from very limited resolution (i.e. 6×10 or 15×10 electrodes) [**2**], [**3**]. Indeed, implanted patients reported having perception of mere light patterns devoid of legibility and needing complex interpretation. Yet, though with low accuracy, this clearly shows that visual activity is indeed recoverable bypassing the eyes.

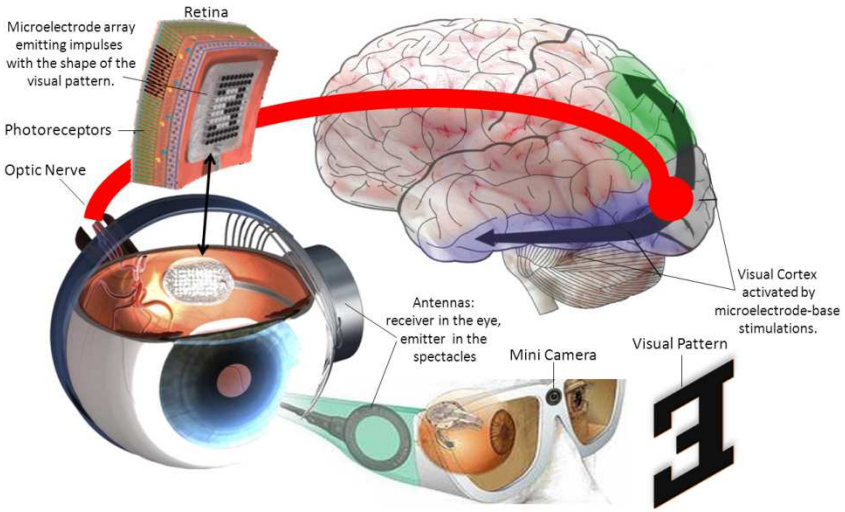


Figure 2-4. An array of microelectrodes is surgically implanted in the retina. Images captured by a camera are processed and converted into electrical impulses by a pocket computer and then, sent wirelessly to the array. The optic nerve gets finally be stimulated with this impulses (therefore the visual cortex too) and a qualia that resembles the image can be experienced. (Modified from wwww.2-sight.eu/)

The blind mind

In congenitally blind things turn out to be quite different. In principle, these individuals have no visual memory and their dreaming seem to be devoid of visual features. Nonetheless, if they suffer from the type of blindness associated to factors other than brain, brain's ability to generate visual experiences must be there in the visual cortex, though unexploded [**33**], [**5**]. Maurice *et al.* [**5**] went further by using PET images to show that when visual information is encoded and convey through another sensory modality, congenitally blind (after training) start to present activity in their visual cortex. However, whether this activity corresponds to actual visual experiences remains largely unknown, as it is the subject of even philosophical debates [**38**] on consciousness. The underlying problem is that structural nature of the perceptual system does not offer any criteria for distinguishing seeing from not seeing [**38**]. Therefore, congenitally blind are not capable of judging quality of visual experiences. For instance, in one of his essays William Molyneux [**38**] posited this question: "would a person blind from birth be able to distinguish visually a cube from a globe upon sudden acquisition of vision?" This is indeed a very complex questioning, since 'normal' vision is itself constrained by top-down knowledge [**14**]. Therefore, sudden acquisition of visual information not necessary implies vision as we know it.

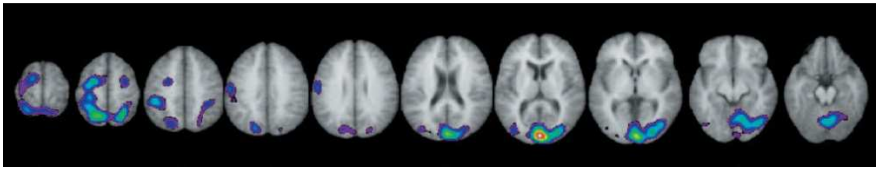


Figure 2-5. (taken from **[5]**)PET data (group analysis) showing visual cortex activations in congenitally blind after trained to receive visual patterns through the tactile sensory pathway (this topic is central to this thesis and will be treated extensively, see for instance Sensory Substitution Devices SDDs (State of the art)). The images refer to from-top transversal slices of the brain. The visual cortex (backside of the brain) is reflected here into the bottom side of the image sequence.

There are bizarre cases of blindness related indeed to the functioning of the brain. Francis Crick and Christof Koch make this point very clear commenting a number of strange cases in "The Quest for Consciousness" **[13]**. Complete cortical blindness ("Mind-blindness" or "visual agnosia" **[39]**) is in fact a consequence of damage to the association cortex of the brain. Particularly Anton's syndrome (from the German neurologist who termed it as blindness of the soul, 'seelenblindheit'), is a rare condition in which patients are blind but deny their condition (they don't know they are blind, nor what seeing means). Although they do not see, they instead have a large repertory of verbal memories with which they confabulate about "visual" things they cannot even imagine **[15]**. Prosopagnosia is another curios case of agnosia in which people experience face-blindness. They are incapable of recognizing faces neither famous nor familiar. All faces look alike to them, so recognizing someone in particular is really hard. They need to adopt strategies like focusing on the voice or any other particular markup of the person. This condition usually leads to social isolation. "The man who mistook his wife for a hat" **[40]** by the neurologist Oliver Sacks, is a remarkable collection of case studies of this sort. Finally, we would like to refer to Akinetopsia **[39]**, **[15]**, a devastating condition of motion blindness. The individual with this disorder lives in a world ruled by strobe lights (like a disco or nightclub). They infer the movement of an object by comparing its relative position in time, though they do not actually see it moving. Others cases of agnosia are color agnosia, depth agnosia etc **[40]**, **[39]**.

What does vision finally entail?

Vision (as the rest of our senses) can be regarded as a phenomenon that entails a twofold task: sensation and perception. Sensation is the function of the low-level biochemical and neurological events that begin with the impinging of a stimulus upon the receptor cells of the eye. In other words, it is the feeling of external visual information as it is registered (sensed) by the eyes. However, visual sensation alone does not imply the coherent conception (or understanding) of external visual objects. 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. In short, we perceive the world through sensations, though we derive sense form it (vision comes into being) only when perception takes place. In this view, Ried [**9**] and Humphrey [**10**] agree that while sensation must be associated with "what is happening to me", perception goes more with "what is happening out there". For instance, "I see a juicy red apple" is a statement that reflects the conception or belief of an external object (perception). This being so, a plausible conjecture is that the eye (and its components) and even the optic nerve are more related to sensation. By contrast, the brain, and more specifically the visual cortex, is more directly likened to perception as it is in charge of decoding visual sensation (through the vast neural network) into meaningful visual imagery.

In the case of visual agnosia the patient has normal sensation although, due to brain damages, perception is never achieved. This is why very often these patients report unawareness of his blind condition: having sensation of "seeing" makes them believe they still can see though, they are not able to adapt those sensations into perceptual experiences. To exemplify this condition we can think of someone speaking in an unknown foreign language. Even though we can perceive the audio cues flawlessly as they arrive to our eardrums (sensation of what is happening to me), we are not able to transform them into coherent information (perception of what is happening out there, or what am I being told). In this order of ideas, we can say that a person who went blind due to damages in the eye or the retina (as it is more common) has lost his sensation skills yet, not the perception ability. In this case, like in visual agnosia, vision of the external world never comes into being inasmuch as the lack of sensation inhibits perception.

However, there is a bizarre condition in which a patient devoid of sensations still has perceptions and therefore vision. Blindsight is the ability of some people to respond to visual stimuli that they do not consciously see. To hold the previous example, let us imagine listening to someone speaking and discovering that we understood his meaning but were unaware of any sound arriving at our ears. In short, a blindsight patient may know that there is an object in front, though he cannot determine where that information came from. Quite opposite to visual agnosia, patients suffering from blindsight deny vision and (by virtue of their sensationless experience) call themselves blind. Curiously, even though blindsight causes lack of sensation, it is not related to any damage in the eye. Blindsight is caused by lesions in their striate cortex and patients suffering from this disorder are known as cortically blind individuals.

As we already mentioned, sensation is a low-level function that has more to do with biochemical and neurological events. Perception being the chief aspect to achieve coherent understanding of the visual world provides a broader framework of study. In fact, most of the functioning of the brain needed for there to be visual meaningful experiences remains unknown. In general, perception involves a lot of more complex information processing than does sensation. Furthermore, perception requires more assumptions and more calculations than does sensation. In recognition of this, O'Regan [**35**] said that visual perception requires

top-down knowledge of the kind of perceptual constancy, for instance. Perceptual constancy is the tendency to conform to the visual object as it is or is assumed to be, rather than as it is presented through the actual stimulus. In other words, it refers to our predisposition to see known objects as having standard size, color, shape, or location regardless of changes in the angle of perspective, distance, or lighting. For instance, when looking at a coin, even though it may appear ellipsoid-shaped because of the perspective, we never assume the coin as being ellipsoid. Rather, we acknowledge its circularity as being affected by relative point of view.

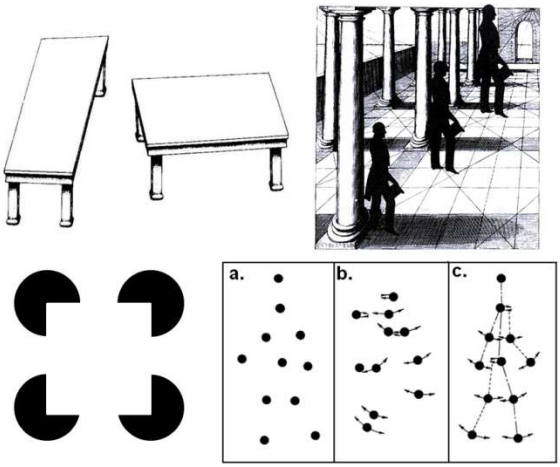


Figure 2-6. Illusions that our brain creates under influence of perceptual constancy.

In **Figure 2-6** we have four remarkable examples of perceptual constancy. On top of this figure we can see two illustrations of how our previous knowledge about perspective laws may affect vision. To the left the tabletop illusion is being presented. It appears that the tabletops have different shapes and sizes but they are amazingly the same. The explanation is that the first table has been drawn as though it was put in perspective. However, none of the expected perspective effects in its shape was actually drawn at all. Yet, our brain distorts the table in haste to derive a coherent view out of the ambiguity presented. In the top-right figure nobody would hesitate to argue that the third man in the rear of the corridor is bigger than the other two. This is because we know that since he is standing in the background, he should appear smaller in proportion to the other two men closer in the foreground. The only way for this not to occur would be that the third man happens to be a giant. And that is exactly what the brain makes believe in order to meet perceptual constancy. Needless saying that the three men in that figure are the same height. On bottom-left of **Figure 2-6** we present the Kanizsa illusion. Again, needless saying that there are only four disks lacking a quarter each. We see, however, a square in between them simply because the visual concept of a square is deeply rooted in our brain. This illusion is so strong (as strong is our visual idea of a square) that the

brain can even complete the missing parts of the square's boundary. Finally, in the bottom-right illustration (**Figure 2-6**) we can observe three possible point-light displays (a, b and c). A static display is not usually seen as representing a human form (a). When the same points light move incoherently or randomly, perception tends to be uncertain (b) [**12**]. However, when the display is in coherent motion, as depicted in (c), it is effortlessly seen as a walking man. Here our brain applies top-down knowledge on kinesthesia and human anatomy to derive sense of the figure.

This top-knowledge provides the kind of information that sighted people get from their visual systems, typically without conscious effort. Kevin O'Regan in his book "*Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*" [**9**] went further on explaining how previous knowledge stored in our brains affect visual experiences. At large, he claims that there are a large number of objective laws characterizing any individual sensory interaction with the world. If we think of red color for example, when we see a red surface the brain codes a subset of the laws that apply, and it registers that they are particular, previously encountered laws. Then, the redness of the red enters our consciousness. Importantly tough, we as persons do not have cognitive access to these laws. In this light, though perhaps in a higher level, we would argue that this is the reason why we do not have to re-apprehend what a face is like every time we see one. This being so, we'd like to conclude this section acknowledging that it takes, for humans to have vision, sensation and also perception. And while the former comes naturally, the latter presents a high dependency on concepts, top-down information, objective laws, perceptual constancy, subjective understanding or to make it shorter, unconscious knowledge. Needleless saying that for there to be vision we rely on physiological and mental factors which, of course, allow sensation and perception of the external visual world, respectively (i.e. the visual system and the functioning of the brain).

2.1.2 Multisensory perception

At any conscious moment we are being "harassed" by sensory information from the external world, and our brains do a remarkable job deriving sense out of it all. It seems easy enough to separate the sounds we hear from the sights we see. Nevertheless, multisensory perception theories reveal that it isn't always the case. In other words, what we see, somehow and somewhat, is always influenced by what we hear, and vice versa. Further, this assumption not only applies to the visual and hearing senses, but to all the reaming sensory modalities. Therefore, how different sensory modalities interact with one another and alter each other's processing, is the focus of research in multimodal integration (or multisensory perception). The central idea here is then, that information from the different sensory modalities is integrated by the nervous system to enable coherent perception of entities, leading to meaningful perceptual experiences [**41**]. In principle, this integration of information is undeniable, at least if restricted to a single sensory modality. For instance, regardless efforts, one cannot see the world in black and white. This is because, in normal circumstances, conscious states

of a sensory modality are highly integrated and cannot be subdivided into components that are experienced independently. Quite in the same way this should extent between different sensory modalities [**42**].

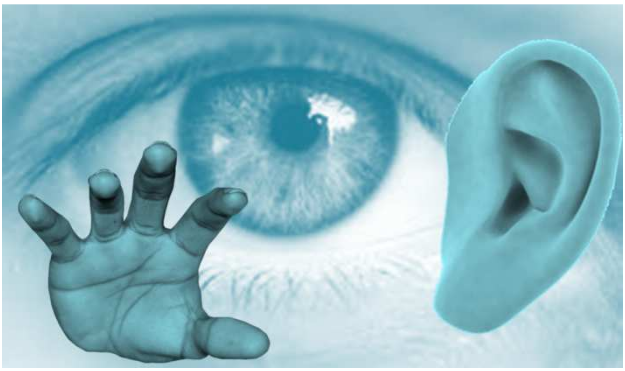


Figure 2-7. Allegorical image on the central idea of multisensory perception or multimodal integration.

To make this point clearer, Koch [**15**] presents a theoretical example: If a person looks at a picture of a car, to say something; neurons in his visual cortex that represent the car's shape will fire while the auditory cortex will stay practically quiescent. In the hypothetical case that all the neurons in the auditory brain could be artificially silenced while the shape neurons will continue to respond to the visual image. One wouldn't be able to hear anything at all. Intuitively, if we suppose there was little sound to hear, that shouldn't make any difference. Yet multisensory perception predicts that even though the brain activity is almost the same (equal sensations) in both cases [**41**], the perceptual experience will differ. This is to say that the fact that neurons could fire but do not naturally, is relevant and very distinct from that in which neurons cannot fire because they have been prevented from.

A practical example to this idea is the very well-known McGurk Effect [**43**]. This brain illusion shows how what we see overwrites what we hear. Specifically, the mouth movements (visual stimuli) of someone speaking can actually influence what one "believes", one is hearing. By contrast, when the eyes are shut, one clearly hears the sound as it is. In other words, if presented with visual evidence of a sound, the brain makes us hear that sound regardless it is not actually the sound we hear. The brain tries to make sense of both cues (visual and audio) so as to reduce ambiguity. This also indicates that the visual sense is predominant compared to hearing when having an integrated conscious perceptual experience [**43**]. Yet both contribute in the subjective experience as such. Similar interaction between hearing and vision in speech perception is given when a movie with deficient sound quality is being

watched. Dialogues in this movie will be mostly understandable provided that the mouths of the actors may be seen clearly [**43**].

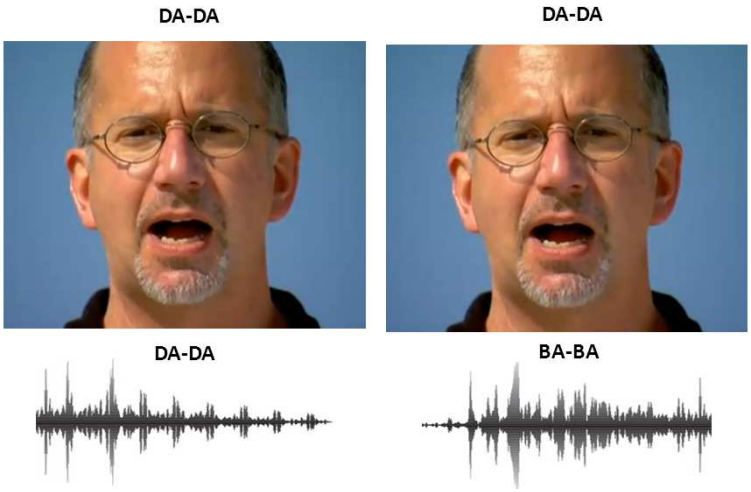


Figure 2-8. The McGurk effect. The visual input is in both cases (left and right) that of a man mouthing DA-DA. As for the accompanying audio cues, in the right case it was deliberately changed for a BA-BA sound, which no longer matches the visual information. The brain refuses to hear the actual sound (BA-BA) and creates the illusion of a DA-DA sound. This illusion is perceivable to any one if viewed in a video[4].

Another aspect of multimodal integration that is worth mentioning refers to the quantity and quality of information that we can derive from perceptual entities, if multisensory perception is purposely promoted during knowledge acquisition [**44**]. This is to say that better information must be attained when several sensory modalities are activated simultaneously [**42**]. For instance, if particular information is heard it will stay in the mind for a certain period of time. If however this information is seen besides; it will be received it in another wavelength, or from another source of input, and it will stay in the head still much better. That's why when attending a lecture, better to have as many visual aids as possible [**44**]. Also writing down notes is recommended because then words are not only listened but seen in the paper too. Besides, the kinesthesia of the hand reinforces learning (integration of three sensory modalities hearing, seeing and kinesthetic movement) [**42**], [**44**]. Although it is not apparent, each sensory modality gives a whole different appreciation of a perceptual object. To give a concrete example, coins can appear circular or elliptical depending on the perspective from which they are viewed. Besides, they might appear smaller when further away and bigger

[4] http://www.youtube.com/watch?v=jtsfidRq2tw

when brought closer to sight. A coin however manipulated haptically in the hand does not appear to distort in either shape or size. Both sensory modalities (seeing and touching) are providing distinct perceptual experiences (qualias) that when integrated are to make up a much robust concept of the coin.

Others have gone further in the line of multimodal integration. At University of Wisconsin, Madison, neuroscientist Guilio Tononi [**16**] says that in fact, consciousness as such is but the holistic result of sensory information being integrated by our brains. He introduced a precise measure capturing the extent of consciousness termed Φ (phi) and expressed in bits. Φ quantifies the information that occurs in a system (e.g. the brain), when it enters a particular state (e.g. a qualia), above and beyond the information generated independently by its parts (e.g. visual cortex, auditory cortex, sensorimotor area, temporal lobe etc.). These parts account for as much independent (nonintegrated) information as possible [**16**]. Thus when taken in isolation, little further integration occurs and the synergy of the system leaves (e.g. unconsciousness). In other words, underlying the unity of a conscious experience there is a multitude of casual interactions among the relevant parts of the brain. In principle, this makes sense, for example, in anesthesia areas of the brain are disconnected and balkanized to the point that consciousness fades [**15**], [**13**] (in terms of Tononi Φ shrinks).

2.1.3 Cross-Modal Transfer and Brain Plasticity

If multimodal integration (multisensory perception) theory is accurate enough; if our brains integrate information from different sensory modalities to create meaningful perceptual experiences; if a sense does not process its stimuli independently, but rather influenced (in some extent) by the rest of the senses. If all this interconnection happens to be true, then it is reasonable to think that a particular sensory perception could be elicited through a sensory pathway that is not typically responsible for it. In practical terms, for instance, if the visual sensory perception is somehow connected (and influenced) to the auditory sensory perception, visual-like experiences could be elicited by hearing and vice versa. This is the central idea in neurological behavior that neuroscientists have termed cross-modal transfer. Furthermore, the neuroplasticity argues that with adaptive learning (i.e. training), perceptual experiences provoked by sensory pathways that are not associated to it, can resemble better and better the actual experience like it was provoked by its typically associated sensory pathway [**33**]. Therefore, back to our previous example, the visual-like experience caused by hearing will resemble an actual visual experience in time. One of the practical effects of neuroplasticity and cross-modal transfers is that if a cortical map is deprived of its input (e.g. blindness) it will become activated at a later time in response to other.

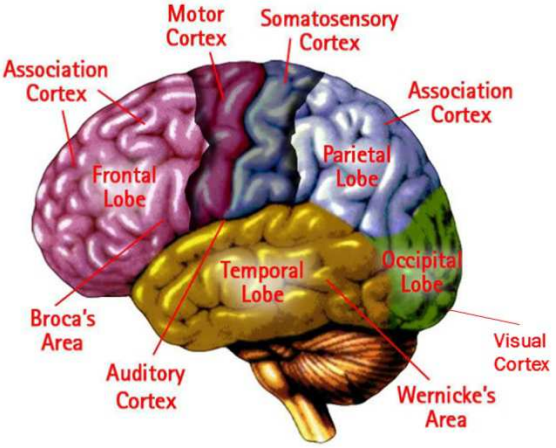


Figure 2-9. Cortical mapping of the brain. Cortical organization in the sensory system is usually described in terms of maps. This is because, as studied in previous sections, visual sensory information, for instance, is almost exclusively reflected into one cortical area (the visual cortex). In sharp contrast, for instance, auditory cues are projected and processed in a different area of the brain. Such somatotopic organization of sensory inputs in the cortex creates cortical representation of the body that resembles a map. In brain plasticity it is believed that remapping of the cortex is possible after, for instance, bodily injuries which promote new conducts in an individual. This idea also implies that if a cortical map is deprived of its input (e.g. blindness) it could become activated at a later time (training) in response to other (i.e. cross-modal transfer).

In 1709 George Berkeley "An Essay towards a New Theory of Vision" [**17**] came to the conclusion that there are no necessary connections between a tactile world and a visual world [**17**]. More recently, in the 90's, Felleman [**45**] and many other neuroscientists ([**46**], [**45**], [**47**], [**48**], [**49**]) showed both, theoretical and experimental support to say that there are no cortical convergence regions, in which neuron clusters integrate information from different sensory modalities (polysensory areas) [**50**]. All of them, from Berkeley to Felleman, maintain the same view that argues against neuroplasticity, they just vary language, after all, more than three centuries have passed. Importantly though, this is in fact what neuroplasticity attempts to change, the formerly-held concept of a brain being a physiologically static organ. This shift in view arises from the believing that changes in neural pathways and synapses are due to changes in behavior, environment and neural processes. And cortical remapping therefore may result from bodily injury (e.g. blindness) which promotes new conducts in an individual, for instance [**50**]. Aside of this debate, one would tend to side neuroplasticity and

cross-modal theories because, for instance, everyone recognizes a key, whether it is felt in a pocket or seen on a table [**51**].

Cross-modal transfer evidence is clear in some particular cases. Synesthesia [**52**], for instance, is a well-documented neurological condition in which stimulation of one sensory or cognitive pathway leads to automatic, involuntary experiences in a second sensory or cognitive pathway [**14**], [**52**]. In other words, in synesthesia stimuli of one sense are accompanied by a perception in another sense. However generally speaking, synesthesia is not considered something that can be learned via training [**52**]. Patients presenting synesthesia-related pathology manifest a broad gamut of dual perceptual experiences: some can taste sounds, others can smell colors, and also in some cases the touch of a texture may trigger a color or even a sound. Importantly, in synesthesia the two perceptions are so vivid that even though one is imaginary, they both seem as though they came from the external world [**52**]. This is to say that even though music may trigger visual perceptions (memory or visual imaginary); the latter is not as sharp as to be considered a synesthesia effect. Nonetheless, while music is a powerful elicitor of subjective emotions, seeing a picture often results as well in subjective emotions. In this regard, Logeswaran *et al.* [**51**] posit the question whether music stimulating the auditory pathway and images doing the same on the visual sensory pathway, both lead to the same perceptual experience (emotions in general)?

Beyond the theoretical aspects, there is of course practical support to cross-modal approaches. For instance, matching two spherical ellipsoids using three different conditions: tactile–tactile (TT), tactile–visual (TV), and visual–visual (VV). Hadjikhani *et al.* [**50**] could identify cortical functional fields involved in the formation of visual and tactile representation of the objects alone and those involved in cross-modal transfer (from vision to touch) of the shapes of the objects. Also, Amedi *et al.* [**53**] show how visuomotor learning affects performance on an audiomotor task, with the aim of proving the cross-sensory transfer of sensorymotor information. More concretely they demonstrated that when a person is exposed to a visuomotor rotation, he tends unconsciously to rotate himself when performing audio guided movements [**53**]. This indicates that the cross- sensory transfer was done naturally [**53**]. Using fMRI images, Tal *et al.* [**54**] identified a network among the occipital, parietal, and prefrontal areas of the brain, showing a clear cross-modal transfer in visual-haptic integration of objects in humans [**54**]. Impressively, Mitchell Tyler *et al.* [**55**] report the case of a female who lost her sense of balance due to antibiotic that destroyed the filaments in her inner ears which transforms sounds into nerve pulses that go to the vestibular system. They designed a "balance device" consisting on a helmet-mounted accelerometer to transmit head and body position to the tongue through electro impulses with and microelectrode array (grid) [**55**], [**5**], [**56**]. The brain of the patient supplied the missing information with that being artificially input on her tongue. She recovered satisfactorily after few months of training with this device [**55**]. In general, cross-modal transfer draws attention in this thesis; therefore more implications of this idea can be seen in (Philosophy and history of Sensory substitution).

2.2 Human cognitive mobility, orientation and space perception

2.2.1 Orientation and Mobility

Mobility is a fundamental task for humans, acquired since childhood. The acquisition of mobility and the guiding principles for this acquisition are however lost to our memories. Mobility is a compound of neurocognitive tasks whose execution varies with the nature of the environment and the data obtained from it. Specifically, mobility encompasses at least three processes. The first, is to understand the near space global geometry. The second, is to walk; i.e. displacement without a specific goal, but with obstacle avoidance. And the third is to navigate; i.e. displacement with a specific goal and with obstacle avoidance. In daily life blind individual perform locomotion (the three tasks) using several assistances. The most popular are the white cane and the guide dog. In this thesis, our purpose is to provide assistance for the first two forms of mobility processes; i.e. understanding the near space and walking. This assistance is tailored to two fundamental human walking strategies, namely the path-integration strategy and the geometry-based strategy [**57**].

Path integration denoted as PI is a continuous process, by which a navigator updates his or her position with respect to a given reference point, generally a point of departure, by processing locomotion signals generated during the physical displacement. PI requires mainly obstacles avoidance, while walking from one to another spatial point. This is the first strategy we learn in childhood, and the only strategy implemented via cane based walking.

Geometry-based walking strategy relies essentially on geometric properties of the navigable space. It is defined by environmental elements relative positions and distances between them, and is supported by mental images of near space, frequently established using city maps and exploratory walking. Guide dogs implement a geometry based strategy in the very near environment. The cane and the guide dog allow very limited exploration of space. Neurocognitive research and the experiences of mobility instructors suggest that the best way of improving on existing mobility aids is to provide in parallel data on all obstacles [**18**] (including their locations, the distances to them and the distances between them). Furthermore,

mobility as a cognitive process, involves several kinds of sensory data, namely touch, vision, balance and hearing.

Human mobility depends on several basic functions of our brain [**58**]. Some of them are: global space perception and understanding (allo-orientation and the anticipation of body movements), self-orientation, walking with obstacles avoidance (without a specific goal), and navigation (walking with a specific goal). Obstacle avoidance requires at least the following functions: obstacle detection and localization, estimation of the distance (and height) to obstacles and estimation of the distance between obstacles.

Mobility therefore involves updating knowledge of one's posture and position using sensory data acquired from the whole environment. These data are memorized in a brain cognitive map [**59**], which drives human actions. Cognitive mapping research focuses on how individuals acquire, learn, store, transform and exploit environment, e.g. encoding locations, their attributes and the relative orientations of landmarks [**60**]. Different elements from the cognitive maps support different human navigation strategies. The most popular human strategies found in the literature on human and mammalian spatial behavior, are path integration, landmark-based strategy and geometry based strategy [**57**].

The cognitive maps formed by the visually impaired differ from those of sighted subjects [**61**]. As a result, their navigation strategies and mechanisms also differ from those of sighted subjects, but these differences have not been sufficiently investigated in current research [**62**]. Mobility instructors teach mainly path integration strategy, as it could be implemented with a white cane based on continuous updates of the end-user position with respect to a fixed point, generally a point of departure [**63**]. The end-users do not have access to global geometry because of the very limited nature of the feedback that can be obtained from a cane.

2.2.2 Space perception

The ability to sense the shape, size, movement, and orientation of objects is known as spatial perception. In perception of spatial relationships sight is the primary sense involved, though other senses such as the hearing also play a role in determining our spatial position within the environment. Processing of spatial perception happens in two levels, first in the sensory organs that gather information from the environment coded into stimuli and then in the brain. Especially, depth perception is a chief element to spatial perception in humans. The brain can approximate the distances between the observer and the observed objects by evaluating their relative size (i.e. perspective). Also perception of the movement (whether objects are moving or still) is important to judge depth. Knowing the relative distance of the objects and the see them in relation to each is also known as spatial awareness and refers to the ability to be aware of oneself in space. Difficulties to acquire proper spatial perception usually lead to conditions such as acrophobia and claustrophobia.

Figure 2-10. Sound-based spatial perception enables blind individuals to play tennis.

In the case of the visually impaired, particularly congenitally blind, spatial perception differs greatly from that of sighted individuals. For the former are limited only to touching distance and very subtle auditory cues, while the latter exploit the visual information which, as already mentioned, is dominant in this task. However, well controlled studies conducted with blind individuals with sufficient experience show that they can function usefully in space. This is to say that vision is important yet not a necessary condition for spatial awareness. A remarkable example to this statement is the blind tennis (**Figure 2-10**). Blind tennis was created in 1984 by Miyoshi Takei [**64**]. Players in this sport use a foam ball filled with metallic that rattle on impact, allowing the blind to locate the ball when it hits the ground or racket. Specialists in this area say that sound localization is so important when blind people navigate the world that it not only can help to practice sport but also with general spatial awareness. Nevertheless, limitations linger in practice, for we know that the capacity of information transfer of the human eye reaches 1000 Kbp whereas, senses of hearing can hardly reach 10 Kbps [**6**].

2.2.3 Stereo Vision

Stereo vision is a technique aimed at inferring depth from two or more close-positioned cameras. This idea is actually inspired on the functioning of the human eyes (**Figure 2-11**). Each eye (about 5 cms of distance between them) captures its own view of the world, so that two slightly different images are sent to the brain for processing. When the two images arrive simultaneously to the visual cortex, they are merged into one. The brain combines these two images by matching up the similarities and adding in the small differences. The small differences between the two images add up to a significant difference in the final image. The fused

image is more than the sum of its parts. It is a three-dimensional stereo image of the world from which we can infer depth and movement of visual objects.

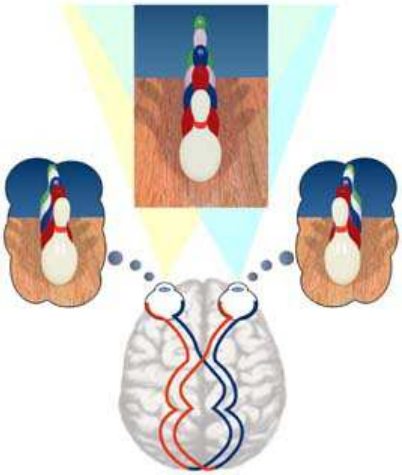


Figure 2-11. The brain perceives depth using a stereo pair of images (one per eye).

Using a camera rig we can infer depth, by means of triangulation (section Efficient registration of range and color images), if we are able to find corresponding (homologous) points in the two images (left and right views of the world). The shift of one point in the left image with respect to the position of its homologues in the right image is known as the disparity. In section Efficient registration of range and color images it will be shown that this disparity is inversely proportional to the depth of the point in the world. Roughly, the closer the points appear to the cameras, the more is the shift thereof in one camera with respect to the other. For instance, one can hold a finger in front of the eyes (at different distances) and view it with each eye in turn to notice how the shift decreases with the distance. Thus, the stereo vision problem (i.e. creating a 3D-image out of a pair of 2D-images) can be regarded as the search of the disparity between corresponding points in the image pair (i.e. the correspondence problem), yet prior constrains need to be considered.

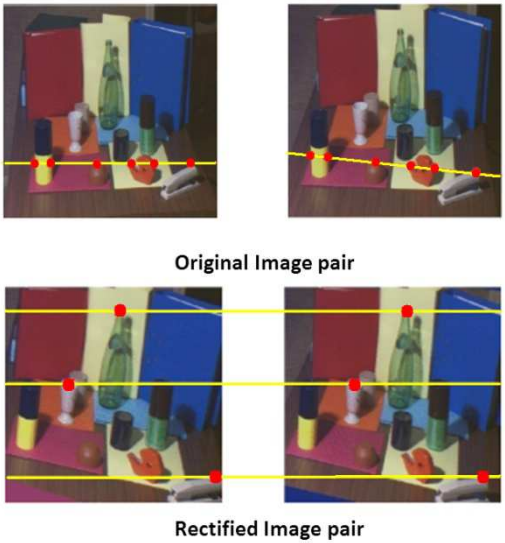


Figure 2-12. The problem of image rectification in a stereo system.

Finding correspondences between points takes a search in two-dimensions for most camera configurations (see **Figure 2-12**). If the cameras are aligned to be coplanar, however, this search is cut down to one dimension (i.e. a horizontal line parallel to the line between the cameras, see figure **Figure 2-12**). Furthermore, if the location of a point in the left image is known, it can be searched for in the right image by searching left of this location along the line, and vice versa. To align the cameras to be coplanar, image rectification is needed. Image rectification is a transformation process used to project two-or-more images onto a common image plane. Also, it corrects image distortion by transforming the image into a standard coordinate system.

Image rectification is possible due to the epipolar constrain. For example, consider two points A and B on the same line of sight of the reference image *R* (**Figure 2-13**). Note that both points project into the same image point $a \equiv b$ on image plane *Pr* of the reference image (**Figure 2-13**). The epipolar constraint states that the correspondence for a point belonging to the line of sight (red) lies on the green line on image plane *Pt* of target image. This image constrain can be made using a linear transformation. A rotation (on x and y axes), sets the images onto a common plane. Also, a scaling transformation makes the image frames be equal and Z rotation turns the image pixel rows directly line up. The rigid alignment of the cameras are known from the calibration [**65**] of the cameras. The calibration coefficients are then used to apply the aforementioned transformations. Finally, the stereo rig can be virtual-

ly put in a more convenient configuration known as the standard form where corresponding points obey the epipolar constrain (blue images in **Figure 2-13**).

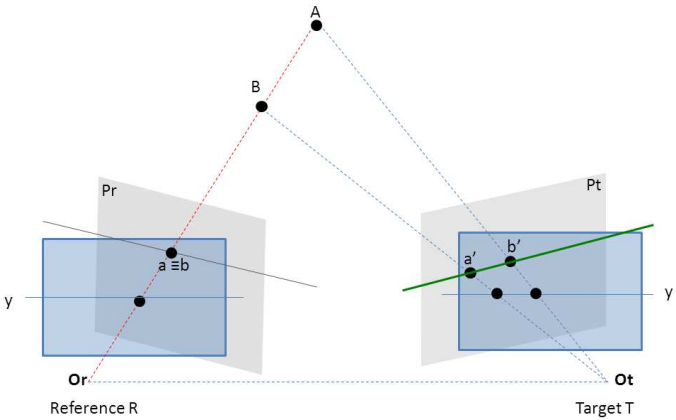


Figure 2-13. The epipolar constrain in a stereo system.

After rectification, the corresponding problem (1-dimensional) can be solved using an algorithm that scans both the left and right images for matching image features. A broad gamut of techniques has been used for this aim. These techniques span from simple normalized correlation (the most popular) up to artificial-neural-networks-based methods:

$$\frac{\sum\sum L(r,c)R(r,c)}{\sqrt{\sum\sum L(r,c)^2 . \sum\sum R(r,c)^2}}$$

Equation 2-1. Normalized cross-correlation of two images.

In Equation **2-1**, L and R represent the left and right images respectively, whereas r and c indicate rows and columns of these images. Note that for finding the matching pair $R(r_j,c_k)$ in the right image of a feature $L(r_j,c_m)$ in the left image, the row j needs not to be vary at all (only the columns $k \neq m$). This reflects the one-dimensionality introduced by the epipolar constrain. To exemplify the stereo vision problem, we can see two images (right 'cyan' and left 'red') of a real scene overlapped in **Figure 2-14** (top-left). Next (top-right), the calculation of the disparity map (using correlation) of the images, has been plotted. Finally in the same **Figure 2-14** (bottom) a 3D image (image with actual depth) based on the disparity is shown. The relation between depth and disparity (section Efficient registration of range and color images) is given by $depth = \frac{Bf}{Disparity}$, where B is the distance between the two cameras and f refers to the focal length.

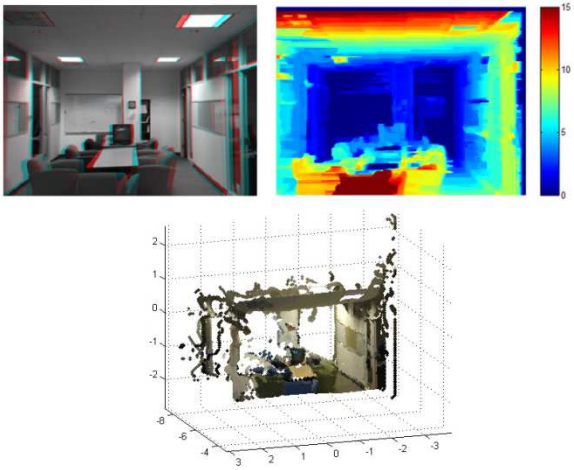


Figure 2-14. 3D image reconstruction using based on a disparity map calculated from a stereo pair of images.

Stereo-vision-based depth calculation, however, gives rise to a number of problems. Specifically, in figure the lack of information in the resulting 3-dimensional image is apparent (sparse maps). This is mostly the case because points (or regions) in the right image may be occluded in the left image (by virtue of the angle difference). This is known as the occlusion problem and yields uncertain disparities (therefore depths) in points whose homologues (in the counterpart image) do not exist (occluded). In this view, the range imaging camera technology has emerged as a method for depth estimation using electromagnetic or similar waves. Particularly, time-of-flight cameras (TOF) are based on a principle which states that the distance between a light source and an object can be deduced with the time of flight given a constant velocity (i.e. time that the light travels to hit the object and go back to the source). In other words, this technology is used in TOF cameras to measure the distance between the camera and each individual pixel in the scene (points). This can be done by either using a laser to scan the scene or an array of lasers (as many as pixels) that project single points. The obtained image is a depth map. More information on range imaging will be given in section Efficient registration of range and color images.

In principle, the chief drawback of these cameras is the lack of color information. However, cost-efficient solutions have emerged recently to alleviate this drawback. The Microsoft Kinect sensor, for instance, is an example of cheap 3D cameras that provide full color [**66**]. This sensor incorporates several advanced sensing hardware. Most notably, it embeds a depth sensor and a color camera. The internal depth sensor calculates object's distances using the structured light principle. It is made up of an infrared (IR) projector and an IR camera. The IR projector is an IR laser that projects a set of IR dots over the object. The relative ge-

ometry between the IR projector and the IR camera as well as the projected IR dot pattern, are all known. Therefore if a dot observed in an IR image is matched with a dot in the projector pattern (using triangulation), it can be reconstructed in 3D. However, the exact technology is not disclosed [**66**]. Much more dense depth maps are achieved by means of range imaging cameras, as shown in **Figure 2-15**.

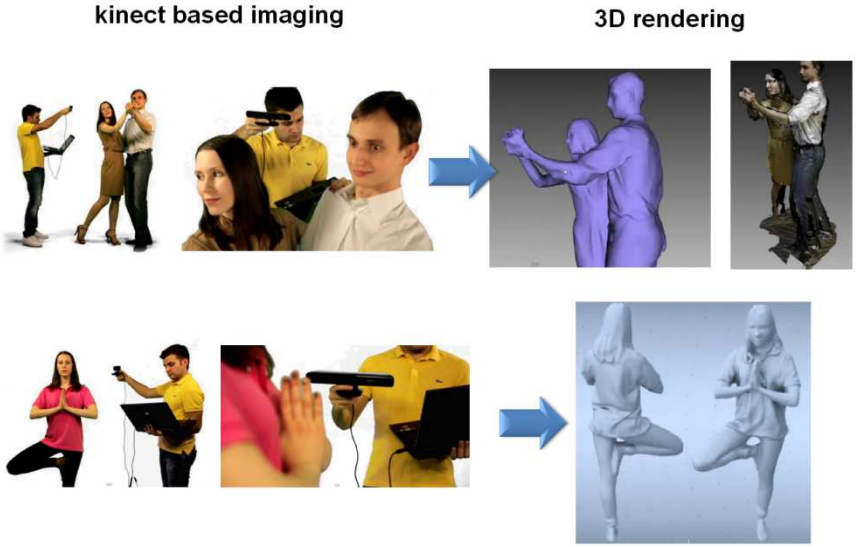


Figure 2-15. Kinect-based 3D images.

2.2.4 Stereo sound

3D sound perception is the ability to locate unambiguously the location of a sound source. For instance, when somebody whispers at one's ear, even with the eyes wide shut, one is able to determine whether the person is located to right or left side of one. A more general example is the sound of the mosquito hovering nearby. One can track the position of the mosquito just by hearing as it flies from one ear to another. In humans this ability to appreciate the auditory space in the sense toward acoustic events is natural, rapid and in general, accurate. It is given mostly by physiological conditions that provide humans with two ears at opposite sides of the head. Sounds enter the brain through two different channels, but once they reach the brain as nerve impulses, a very complex mixing takes place. Depending on the position of the emitting source, some impulses reach one side of the brain; others reach the opposite side and some irradiate both sides. This is what finally permits us to hear in "three dimensions" (left, right and depth) in a quite similar way we have depth perception with two eyes (as seen

before). Just by having a listen to outdoors with eyes closed, we can experience this phenomenon. One can identify birds chirping and furthermore, one can judge how far away they are based on hearing them only. In other words, a 3D audio image of the environment (also known as soundscape) is traceable. Directional and 3D hearing is not possible with only one ear.

Humans locate 3D sound due to two different cues (delay and strength of the audio). For instance, let us suppose a sound source irradiating from the right side. The time that takes for the sound to get to the right ear is slightly shorter that the time that takes for the sound to reach the left ear. The time difference between the two hears is enough for the brain to realize where the sound is coming from. The second kind of cue is that sound as it travels from the source, when it hits the right ear it has a certain level and by the time it hits the left ear it has a lower level. These cues encode the source location, and may be captured via an impulse response which relates the source location and the ear location. This impulse response is termed the head-related impulse response (HRIR). HRIRs can be recorded using two microphones placed inside the ears of a dummy head as shown in **Figure 2-16**.

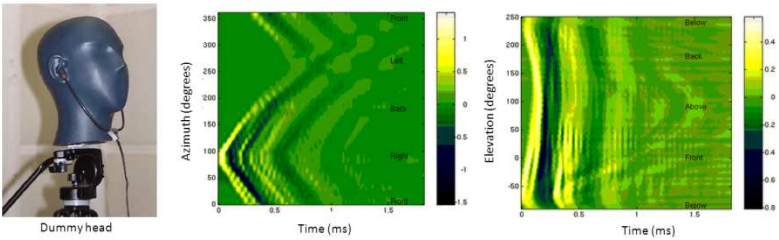


Figure 2-16. This figure shows experimentally measurements of head-related impulse response (HRIRs) using a dummy head (with microphones in both ears). The graphics show the response of the right ear to an impulsive source as moved through the horizontal plane (middle graphic) and elevation (right graphic). The strength of the response (level) is represented by brightness. For instance, we can see that the sound is strongest and arrives soonest when it is coming from the right side of the azimuth plane (middle graphic). By contrast, when the source is moved top-down around the head, the changes are subtler. Arrival time is pretty much the same, as one would expect (right graphic).

It is worth noticing that we do not hear in three dimensions just by using a pair of speakers (right and left). This is because even though the left speaker contains the 3D cues for the left ear, these cues get corrupted when the left ear hears the right speaker and vice versa (i.e. cross talk). Without canceling the cross talk the cues are unclear and the brain won't get the information it needs to hear in 3D. Essentially putting a virtual wall between the two speakers is the key to turn regular sound into 3D audio. This can be done using a filter that roughly sends negative and positive pressure waves over each speaker (compensation). Moreover, this filter achieves the presentation of externalized spatial virtual sources (i.e. providing the

illusion of sounds originating from specific locations in space), for it models the time and the level of the sound as it reaches both ears. In the literature this is referred as a Head Related Transfer Function (HRTF) [**67**] [**68**].

Essentially, the HRTF is the Fourier transform of the HRIR and captures all of the physical cues to source localization. Once the HRTF for the left ear and the right ear are known, one can synthesize accurate binaural signals from a monaural source. In other words, the HRFT describes how a sound from a specific point will arrive at the ear. Consequently, the transfer functions to the contralateral ears (H_{LR} and H_{RL} in **Figure 2-17**) need to be zero (cross talk canceling) while the transfer functions for the direct transfer paths need to be equal to one (H_{LL} and H_{RR} in **Figure 2-17**). This principal can be formulated in matrix-vector notation. For instance in the case of two sources, the source output signals Y_{right} and Y_{left} can be calculated from the input signals (in the ears) X_{right} and X_{left} as follows:

$$\begin{bmatrix} Y_{left} \\ Y_{right} \end{bmatrix} = H^{-1} \begin{bmatrix} X_{left} \\ X_{right} \end{bmatrix},$$

$$H = \begin{pmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{pmatrix},$$

Equation 2-2. Matrix form to estimate the HRTFs of two sources.

with H^{-1} being the inverse of the matrix H that contains all the transfer functions shown in **Figure 2-17**.

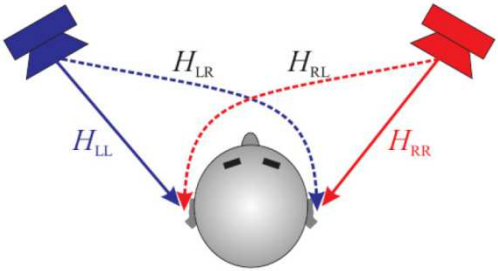


Figure 2-17. Transfer functions for two sound sources.

Finally, it is seen in **Figure 2-16** (right) that the distinction of sound sources in elevation turns out more difficult as sounds arrive to the ears almost equally (time and level). This is mostly the case because our sound receptors (the eardrums) are left-right located. In fact, this is a very subjective perception hard to be represented by mathematical models, so that it often requires personalized HRTF measurements as shown in **Figure 2-18**. This is why, rather than 3D audio, many applications use just 2D audio (i.e. spatialization through the

azimuth plane). This kind of spatialization creates the illusion of sound coming from virtual sources placed from left to right (not elevation). To achieve this effect we do not need to calculate the HRTF (which is a complicated function of four variables: three space coordinates and frequency). Just the convolution of an arbitrary source sound with standard HRIR converts the sound to that which would have been heard by the listener if it had been played at the source location, with the listener's ear at the receiver location. In fact this is the method that we will use in this thesis and it will be further discussed in section Lateralization of sound.

Figure 2-18. Taken from [69]. Equipment for fast personalized HRTF measurements. The speakers located on the arc produce wide spectrum sounds recorded by microphones placed in the ear canals. The chair rotates automatically, powered by a step motor, allowing for fast collection of HRTF data from all directions around the listener.

2.3 Sensory Substitution Devices SDDs (State of the art)

2.3.1 Grounds and history of Sensory substitution

Sensory substitution refers to the mapping of stimuli of one sensory modality into stimuli of another sensory modality. This is usually done with the aim of bypassing a defective sense, so that associated stimuli may flow through a functioning sense. This benefits the handicapped people to the extent that some loss of functioning in a sense is restored. It was already studied in previous sections (Blindness) that when individuals go blind o deaf, they do not actually lose the ability of seeing or hearing, rather they become incapable of convening external stimuli to the brain. Since 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**]. This idea largely relies upon the concept of brain plasticity [**70**] explained also earlier in this thesis. Roughly, brain plasticity denotes the self-adaptation ability of the brain to the deterioration (or absence) of a sense [**70**]. This is a natural mechanism that allows people devoid of a sense to adapt and compensate through other sensory pathways. For instance, cortical remapping or reorganization happens when the brain is subject to some type of deterioration [**71**].

Sensory substitution may be achieved by using invasive methods that collect external signals and transduce them into electrical impulses for the brain to interpret them naturally [**38**]. Thus, stimulation of the brain without intact sensory organs to relay the information is possible [**34**], [**70**], [**71**]. However, in this section and in general in this thesis, we aim at exploring the broad gamut of non-invasive sensory substitution devices also known as SSDs. These devices use human-computer interfaces to transmit sensory information (of a substituted sense) captured by an artificial modality (artificial sensor) to another human sense (substituting sense). In other words, they translate sensory information in order to enable perception through another than the originally responsible sense [**4**]. Nonetheless, detailed information about the formal definition of these devices will be given latter in this section.

Kercel *et al.* [**38**] consider reading to be the oldest sensory substitution system because acoustic information (spoken words) is presented in a visual way (writing) [**38**]. According to this view, for this particular case the bridge between the two senses is made through ink and paper. In similar line, the Braille system, developed in 1840 by Louis Barille with the purpose to aid the blind acquiring visual information through touch, could be considered to be the most popular sensory substitution method [**38**]. Later in 1897, Kazimierz Noiszewski prototyped the first technical sensory substitution device termed Elektroftalm [**72**]. The idea was simple, in order to enable the blind to be aware of dark and light spaces, he used a light sensitive selenium cell. Thus, brightness was encoded into audio cues that aided the subjects to distinguish the binary situation. Though this idea was evolved in time, for instance as head worn system (see **Figure 2-19**). Most of the literature agrees that the first formal SDD did not appear until several decades later in the 20th century.

Figure 2-19. (taken from [73]) Later version of Noiszewski's Elektroftalm (1970).

The first SSD broadly accepted by the scientific community goes back to the late 1960s. Paul Bach-y-Rita (an American neuroscientist whose most notable work was in the field of neuroplasticity) introduced a preliminary prototype device, based on the tactile sense [**74**], to gain and relay environmental information mainly to the sense of vision [**39**], [**75**] (see **Figure 2-20**). This SSD would be known as the Tactile Visual Sensory Substitution (TVSS). In his worldwide-known book "*A History of the Mind: Evolution and the Birth of Consciousness*", Nicholas Humphrey [**10**] describes Bach-y-Rita's first trials with human subjects at the Smith Kettlewell Institute: A participant was provided with a black-and-white TV camera (attached to his head), whose electronic image, was sent to a matrix of vibrators attached to the back of the subject in contact with the skin. This matrix had 400 vibrators arranged within a 20 x 20 grid, covering a 10-inch-square area of skin (see **Figure 2-20**). Thus each point stimulated on the skin represented one pixel of the image captured by the camera, as shown in **Figure 2-21** (In the original text they use the term point rather than 'pixel', as it was very unfamiliar back in those days). A participant could direct the camera by moving his head, somewhat as moving his/her own eyes.

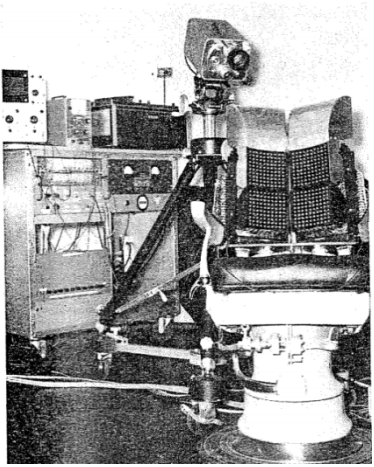


Figure 2-20. (taken from [74]) Tactile television hardware (1969) comprising the vision substitution system. The digitally sampled television camera with zoom lens is seen high in the center; the electronic commutator and control electronics with monitor oscilloscope and videotape recorder are on the left. On the right, the 400 point two-dimensional tactile stimulator matrix array is shown mounted in the back of a dental chair for projecting mechanical television images on to the skin of the back of blind subjects. In the position shown, the camera permits subjects to examine hand held objects from a visual angle approximating that the eyes. When placed in front of the subject the camera can be manipulated to examine various parts of an object. [74]

The results surpassed all sort of expectations, with few hours of training, visually impaired individuals learned to identify a variety of objects such as a toy horse, a cup and a telephone. According to Bach-y-Rita reports [**76**], subjects rapidly gained skill to point accurately to objects, and to estimate their distances and sizes. Furthermore, with about thirty hours of training participants were enabled to perform complex pattern discriminations and, to everyone's amazement, various subjects could recognize the faces (**Figure 2-21**) of some persons: "*That is Betty; she is wearing her hair down today and does not have her glasses on; her mouth is open, and she is moving her right hand from left side to the back of her head*" [**76**]. Consequently, Bach-y-Rita did not hesitate to saying that these blind subjects had regained some visual perception: "*If a subject without functioning eyes can perceive detailed information in space, correctly localize it subjectively, and respond to it in a manner comparable to the response of a normally sighted person, I feel justified in applying the term 'vision'*" [**76**]. Later in the literature, this would be known as the behavioral criterion to evaluate SSDs performance [**14**].

Figure 2-21. (taken from [74]) Appearance of a 400 point representation of a woman's face as seen on the monitor oscilloscope (1969). Subjects can correctly identify vibrotactile stimulus patterns of this level of complexity. Blurring and consequent half-tone appearance in the image occurs visually (and tactually) due to noise modulation and temporal integration of the 60 Hz field rate. (Visual perception of this type of digital display in sometimes enhanced by squinting or otherwise further blurring the image).

In **Figure 2-22** we present a graph model that globally describes Bach-y-Rita's experiment. The aim of this figure is to comprehend Bach-y-Rita's idea of proposing sensory substitution as a kind of acquired synesthesia[5] [**14**], [**77**]. This idea enjoys wide acceptance in recent research of SSDs [**71**], [**14**], [**5**], [**70**], [**34**]. Note that a stimulation of one sensory pathway leads to automatic, involuntary experiences in a second sensory or cognitive pathway [**77**]. In other words, while tactile sensation remains in the stimulated subject, this very stimulation elicits a visual-like sensation not related at all with touch. Since vision in this experiment is not seen as estimation derived from touch, but a visual experience as such; Bach-y-Rita coined the term skin-vision. **Figure 2-22** could be regarded as well as the earliest framework for a visual sensory substitution system. Due to such a remarkable experiment, sensory substitution devices became the basis of multiple studies investigating perceptive and cognitive neuroscience, computer science, electronics, and more recently, human computer interaction HCI.

[5] A sensation experienced in a part of the body other than the part stimulated

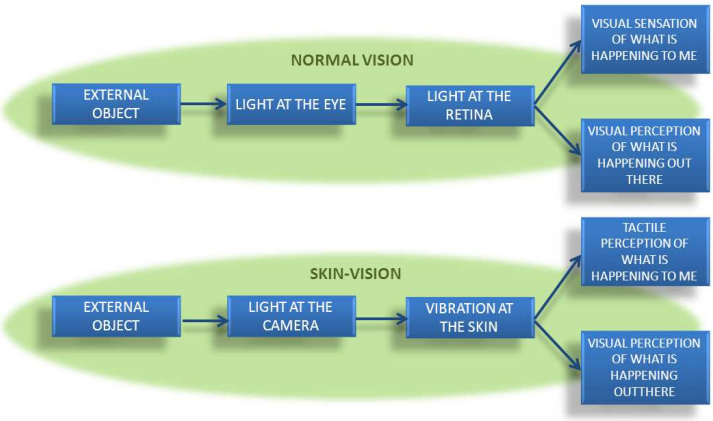


Figure 2-22. Two contrasting situations: the case of normal vision, and the case of skin-vision.

There is no reason to limit SSDs to visual substitution, though non-visual SSDs are less common. In this view, **Figure 2-23** shows all possible sensory substitutions that give rise to a variety of potential devices. In order to generalize the idea of SSDs, however, various authors [**72**], [**4**] have agreed in an overall structure (**Figure 2-24**) to model sensory substitution devices (regardless the substitution wanted). At large, a sensory substitution system can be thought as a composition of elements of the kind illustrated in **Figure 2-24**: A sensor, a coupling device and the actuators. The sensor is a device that captures information *x(t)* from the environment so as to feed the coupling device. This sensor may present one (modality A) out of two, namely: only-receiving (e.g. camera, microphone) or receiving/emitting (e.g. laser, ultrasound). More importantly, this sensor needs to be autonomous handling by the user, so that the modality under which the user interacts with the sensor is also accounted in **Figure 2-24** (Modality B, more often motion). The actuators are made up of a user and the display of the coupling device. This latter deepens on the type of sensory substitution system (e.g. headphones, tactile display) and it is in charge of the stimulation of the substituting sense that, in turn, forwards information to the brain. Then, the body reacts and makes the sensor to move in pursuit of new data *x(t)*. As for the coupling device, this is the one that bridges the sensor and the display by transforming *x(t)* into *y(t)*. The data *y(t)*, therefore, is seen as stimuli *x(t)* of the substituted sense mapped into stimuli of the substituting sense.

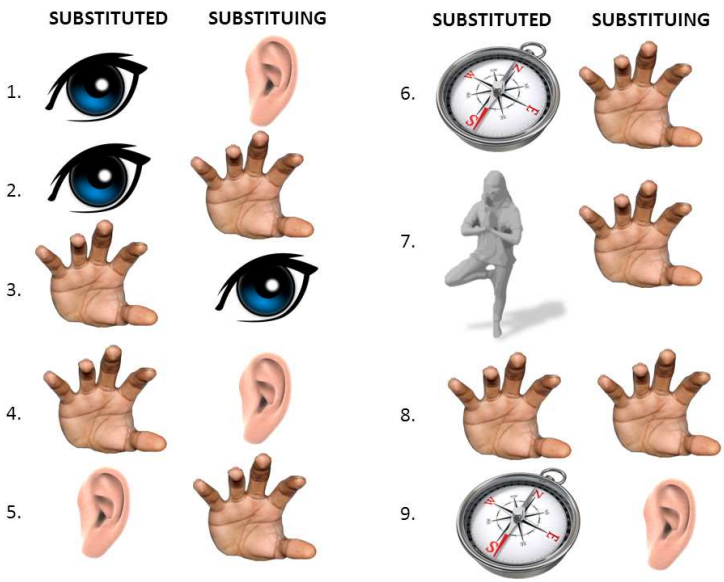


Figure 2-23. Possible sensory substitutions (from 1 to 9): Audible Vision, Tactile Vision, Visible Touch, Audible Touch, Tactile Hearing, Tactile Orientation and Spatial Awareness, Tactile Balance, Tactile sensory relocation, Audible Orientation and Spatial Awareness. Since the subject of this thesis is visual substitution, the two modalities aimed at substituting this sense (i.e. audible vision and tactile vision) are reviewed in this state-of-the-art section. Moreover, since See ColOr is not intended for tactile vision, it might be classed as an audible SSD, though not necessarily limited to. It will be studied later in this thesis that See ColOr may involve tactile sensory relocation (kinesthesia) in one of its modules of interaction.

In the particular case of visual substitution, the structure shown in **Figure 2-24** may be synthetized as consisting of: a visual sensors (i.e. camera) that relay information to a coupling device (e.g. a laptop, or a Field-programmable gate array –FPGA–) that systematically translates visual features into tactile or auditory stimuli outputted finally, through headphones or mechanical/electrical haptic stimulation devices. It is worth noticing here that given the technical difficulties related to the generation of tactile stimuli, the proliferation of SSDs for the visually impaired have largely targeted audition ([**78**]; [**79**]; [**80**]). Nevertheless throughout this section both modalities will be revised Tactile Visual substitution and Auditory Visual substitution, though the former in not much detail.

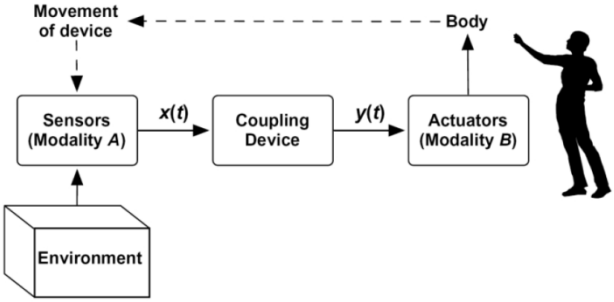


Figure 2-24. (taken from [72]) Structure of a sensory substitution system.

Finally, it is important to stress the fact that there may be certain devices closely similar to SSDs, though they do not fully meet the definition of sensory substitution [**14**]. Rather, they may be regarded as sensory augmentation or extension devices (see General Discussion). In particular, for visual substitution, the white canes serve as example. Although they are intended to restore some normal functioning for the blind, they extend the physical range of a functioning senses (i.e. touch), instead of substituting the visual sensory pathway via a coupling system [**14**]. In fact, it is very commonplace observation that many visually impaired individuals report feeling their white cane or their dog as an extension of their bodies.

2.3.2 Tactile Visual substitution

Although the scope of this thesis does not span tactile substitution of vision using SSDs,we would like to make reference to some of the most relevant examples in this field. However, other SSDs specifically intended for color substitution through the sense of touch will be cited in Sensorial color coding section.

Optacon

The Optacon is a tactile sensory substitution device developed by Bliss *et al.* [**81**] in the early 60s. This SSD was designed as a tactile-based reading aid. Nonetheless, it was latter enabled (with the addition of a lens with a more distant focal plane [**81**]) to provide visual environmental information. The tactile display of the Optacon is made up of a 6x24 pin array designed for the fingertip. TeleSensory [**82**] was a commercial version of the Optacon that was available for sale until mid-90s, when the device went out of production [**82**]. The decline in popularity of the Optacon was marked by the apparition of optical character recognition OCR techniques and eventually computer-based screen readers. The reason being, these tools required much less training to use. However, the Optacon continued to be used in many

studies on sensory substitution and tactile perception. An illustration of the system and its functionality are shown in **Figure 2-25**.

Figure 2-25. (taken from [**34**]) Using the Optacon the child in this image was able to mimic the hand posture of the teacher using feedback acquired via the device. The first prototype was completed on 1969. It was portable and combined the stimulator array, electronics, batteries, and camera in a single package measuring 13.5" by 8" by 2.25". The total weight of this device was 9 pounds.

Tactile Artificial sight

Back in 1982, Thierry Pun at the "Ecole Polytechnique Federale de Lausanne (EPFL)", presented his PhD thesis [**83**] that bore the name of: Automatic simplification of scenes using image processing for tactile restitution of sight handicapped. This work aimed at designing an image processing unit to generate simplified tactile version of images. Hence, an efficient picture simplification scheme for tactile outputs was proposed [**84**]. By and large, this visual-to-tactile SSD idea introduced two significant contributions: firstly, the psychological considerations of tactile recognition mechanisms within the conception of methods for image processing. Secondly, a new method for grey-level picture thresholding using the entropy of the histogram [**85**]. One of the possible applications of this work was the development tactile images for educational purposes. Pun thought that this could be used in classes for blind children, providing relief illustrations from grey-level documents. Quite unlike the Optacon though, this approach was much more concerned with the optimal processing of the images in order to yield best tactile outputs in terms of understandability. As a consequence, the blind children would be allowed to develop their mental images easier, faster and more accurately. This SSD was thought as an "on-line" portable device that would be based on relief printers, yielding 0.2 to 0.6 mm high outputs. Finally, this project also included the creation of synthetic textures to ease in the tactile discrimination of regions.

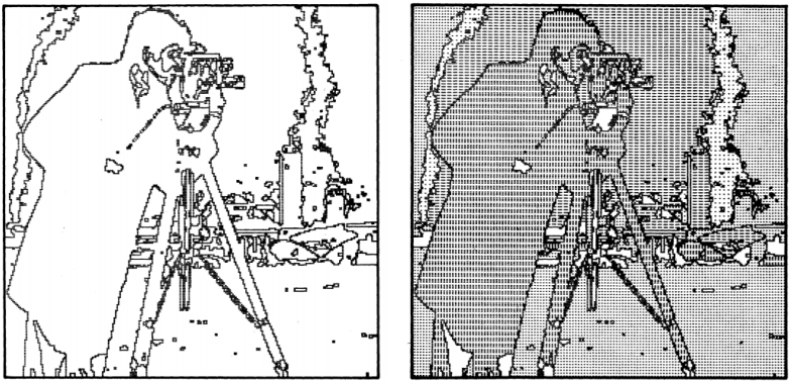


Figure 2-26. (taken from [84]) Tactile Artificial sight's output images: Right, edge-based image. Left, introduction of synthetic textures.

TDU (Tongue Display Unity)

This device displays images on the dorsum of the tongue using a flexible small electrotactile array of electrodes. Over time, The Tongue Display Unit or TDU [**56**] (which we also identify as a BrainPort because of the similarity to a computer USB port), trains the user's brain to translate tactile to visual information so as to enable him to see the images captured by a head-mounted camera. A computer converts the image (gray level) to pulse trains that are then carried to the electrode array (on the tonge) via the ribbon cable. Trial subjects report to experience the resulting stream of sensation as an image. In particular, the tongue is very sensitive because the sensory receptors are close to the surface skin. This SSD has been developed to take advantage of this characteristic. In fact, some authors have shown [**56**] that form perception is somewhat better achieved with electrotactile stimulation of the tongue rather than the fingertip. Besides, the tongue needs only 3% of the voltage (5–15 V), and much less current (0.4–2.0 mA), than the finger pad [**56**]. The pulse trains are carried to the electrode array via the ribbon cable, and the electrodes stimulate touch sensors on the dorsum of the tongue. The subject experiences the resulting stream of sensation as an image. An example of the use of TDU can be seen in **Figure 2-27**. This system enjoys great acceptance in terms of usability and is still commercialized. Furthermore, the example described in the last paragraph of section Multisensory perception provides a clear example of the multi-usability of this system. In that case it was successfully used for subtitling the vestibular sense in a real patient.

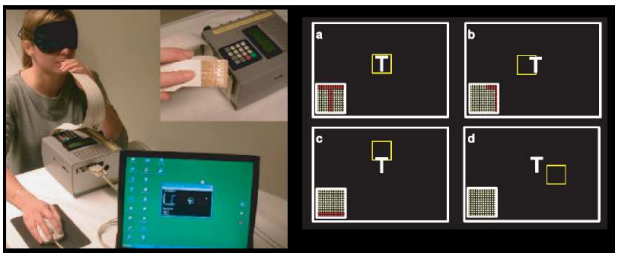


Figure 2-27. (taken from [**5**]) A blindfolded sighted individual using the TDU to recognize the orientation of a letter. Note that for this experiment in particular, the area of the image being transmitted to the tongue (in yellow) is moved via mouse.

Color perception

SmartTouch uses optical sensors to gather visual information and electrical stimulation to translate it into tactile display. Tachi *et al.* [**86**] developed this electrotactile display to enabling the composition of tactile sensations by means of selective electrical stimulation of tactile mechanoreceptors. The authors described such phenomenon as "tactile color". This work is based on the study of the electrical physics of the skin in relation to the geometry and composition of its afferent neurons [**86**]. At large, when the display (attached to the fingertip) contacts an object, tactile sensations are said to be felt as vibration or pressure. Thus, an individual not only makes physical contact with an object, but also touches the surface information (i.e. color captured by the optical sensor). This is known as augmented haptics, derived from the concept of augmented reality: a perception of the real-world in which elements are added (the reality is augmented) via computerized techniques. A schema of this idea can be seen in **Figure 2-28**.

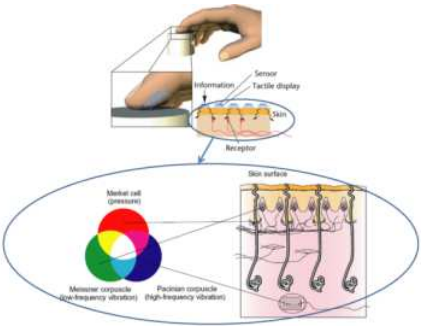


Figure 2-28. (taken from [86]) Smart Touch's color encoding through tactile stimulation.

2.3.3 Auditory Visual substitution

Here we present a chronological review among the most relevant events (experiments) in the literature that have shaped the evolution of vision-to-audio SSDs, as we know them today. In fact, we make reference here in this section to both the most cited works in SSDs' literature, as well as those that for the best of our knowledge we consider worth comparing. Importantly, as great part of this thesis is concerned to the representation of color in SSDs, here we mention some relevant works in this particular topic, though further discussion and more approaches are given in section Sensorial color coding. We will conclude this section mentioning at least three significant systems within the category of "augmentation of real objects with virtual sounds". Although these approaches show a family resemblance to SSDs, they do not meet the more restricted definition of sensory substitution. Rather, they are new technologies intended to discriminate, synthetize and communicate information of the world to the blind. Indeed, they are not used in rehabilitation but mere assistance, given that no neuroplasticity or visual skills regaining are intended. Nevertheless, such approaches turn out attractive to this work, since our final goal is to combine both, typical SSDs and this sort of systems that augment real objects with virtual sounds, particularly those based on computer vision.

An audio display for the blind (1975)

In 1975, RAYMOND M. FISH wrote an article [**87**] to introduce an auditory code capable of transducing two dimensional patterns and pictures for presentation to blind people. He used a television camera to acquire input images and a series of electronic circuits as to translate them into code. To encode a picture, a sequence of tone bursts were used representing the black dots that shape the image (i.e. pixels). The vertical location of a pixel was revealed to the blind individual (wearing binaural headphones) by the frequency of the tone burst. The horizontal position of the pixel, in turn, was represented by the ratio of sound amplitudes presented to each ear. Note that this is but an early attempt of spatialized sound to create the illusion of source emitting from specific location from left to right. In this article the author claimed: "These methods of coding vertical and horizontal position utilize the principles of psychoacoustics in that high pitched tones naturally seem to come from a high location, and amplitude differences in the sound presented to the ears make it seem that sounds are coming from certain positions from left to right" [**87**]. Finally, this 1975 paper [**87**] reported that both blind and blindfolded subjects could use the display coupled to a TV camera as a mobility aid to travel within simple indoor environments with just 20 minutes of training. Further, a ten-year-old blind subjects with 4 hours of training, was able to identify patterns as complex as those shown in **Figure 2-29**.

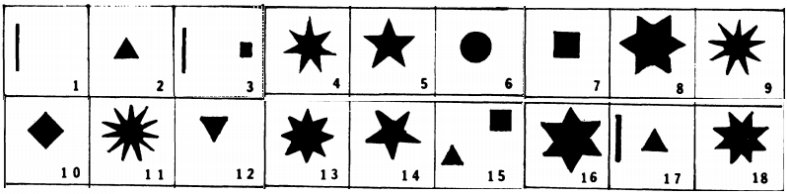


Figure 2-29. (taken from [87]) Patterns recognized by 4-hours-trained blind subjects in this 1975 experiment.

Blind babies see with the ears (1977)

Figure 2-30. 1977-cover-page of Newscientist magazine: "Sooner or later technology will rescue blind people from their prison of sightlessness. Rapid advances in electronics are even now making possible aids and prostheses thought impossible a decade ago. Yet as these progress reports on two promising approaches conclude, the factor limiting advancement is no longer technology but inadequate understanding of the psychology and physiology of perception". [88]

Here the concept of sensory substitution using sound to compensate for blindness was applied to babies, taking advantage of their early development. Development is an accumulative process in which early acquisitions determine the possible direction of later events processes [88]. So the assumption underlying is that babies will acquire the ability to see through sound nearly naturally, if taught from their earliest ages. This device provides the babies (in these

experiments) with sound information about their environment using ultrasonic echolocation similar to that of a bat. The SSD constantly irradiates from objects into audible sound. The adaptation from ultrasound to audible sound codes the echo in three ways: "The pitch of the audible signal is arranged to indicate the distance of the object from which the echo came. High pitch means distant objects, low pitch near ones. The amplitude of the signal codes for the size of the irradiated object (loud-large, soft-small) and texture of the object is represented by clarity of the signal. In addition, the audible signal is 'stereo' so direction to the object is perceived by the difference in time of arrival of a signal at the two ears" [**88**]. Here again as in [**87**] , we find preliminary approaches to spatialized sound. Finally, this whole experiment is summarized in **Figure 2-31**.

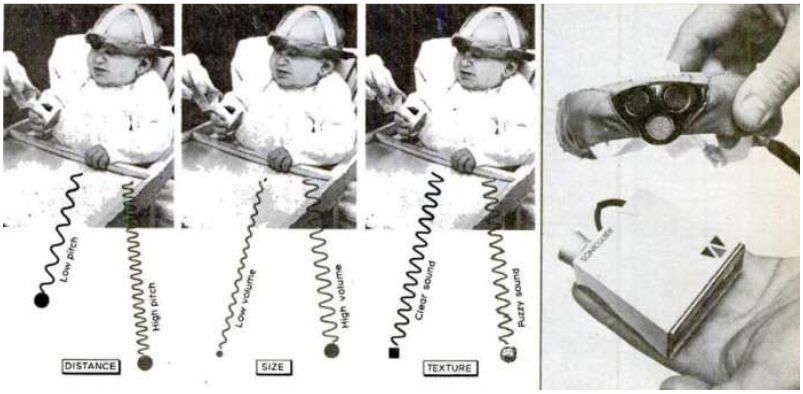


Figure 2-31. (taken from [88]) How a baby can recognize an object using ultrasonic echo location.

It is worth noticing that in this work the novelty was the recruitment of congenitally blind babies, since echo location based on ultrasounds had been previously used in blind assistance. As an example, "The Sonic Guide" [**89**] (**Figure 2-32**) was created in 1966 as a frequency-modulation-based ultrasonic aid. A single transmitter is mounted centrally on the bridge of the spectacles. Receivers are mounted on each side of the transmitter, one directed to the left and one to the right. The echo signals collected by the receivers are transposed into the audible band and fed to a stereophonic display. The splay angle between the receivers is critical and is chosen so that the stereophonic sound image produced by the presence of an object in the beam has an azimuth displacement identical to that of the real world object causing the echo [**90**].

In fact, in the 80's there was a proliferation of these devices with evolved characterizes, known as ultrasonic pathfinders [**90**]. Nowadays, we can find modern versions of these SSDs sold in the market as audible electronic mobility aids designed for individuals who are blind

or have low vision (**Figure 2-32**). At large, these electronic obstacle detectors use a head-mounted pulse-echo sonar system controlled by a microcomputer, giving the visually impaired advanced warning of objects which lie within his or her travel path.

Figure 2-32. The evolution of so-called sonic pathfinders from the 60's (left [89]) up to 2012 (right). In the 80's there was a proliferation of research and production of such devices.

The vOICe (1992)

The vOICe (by Peter Meijer, 1992) is a sensory substitution device, which sonifies 2D grayscale images, allowing its user to "see with sound" [**78**]. As noted in Philosophy and history of Sensory substitution and as it will be discussed later in General Discussion, here the use of the word "see" (same as for Bach-y-Rita [**31**]) has more implications than a mere analogy. 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). The images are coded in horizontal, one second sweeps, so that the horizontal position of a column of pixels is represented through the time position in the sweep. The vertical coordinates are assigned different sound frequencies and the amplitude of the frequency component corresponds to a given pixel's brightness. This coding method creates a one second audio signal, with numerous frequency components and a time-varying spectrum (**Figure 2-33**). The sound signal coded by the vOICe can be easily decoded by a machine (**Figure 2-34**); however, the question arises whether such a complex solution can be functional for human beings [**78**].

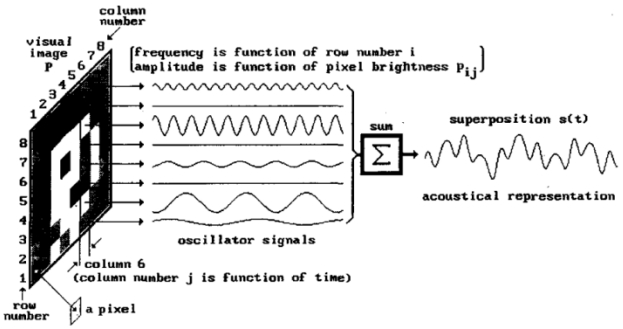


Figure 2-33. (taken from [78])The vOICe visual-to-audio encoding schema.

It has been proven, that after prolonged use and with correct training, a user's brain will actually adapt itself to the new perception method [**91**]. After initial training a user can distinguish various artificial images or simple objects on contrasting backgrounds; however, to use the skills in real world scenes requires months of practice and even more. After extensive training with this SSD, the user may be able to perform feats which seem impossible for a blind person, such as picking up a cup of coffee from a table without having to tactilely probe the environment. The trade-off however, is that the signal from the vOICe requires a lot of concentration from the user, blocks out a lot of environment sounds, and can be very irritating over time.

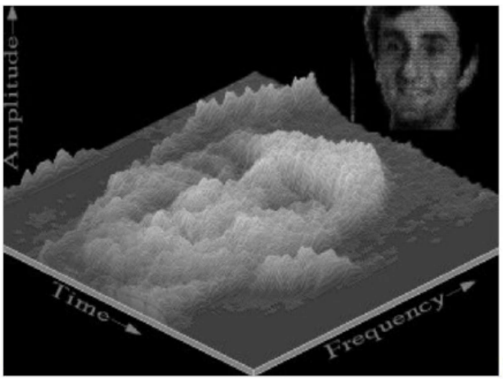


Figure 2-34. (taken from [**78**]) The sound spectrum of a sonified image using the vOICe.

An advantage of the vOICe is that it can be applied to other tasks aside from mobility. In time, after the brain adapts to the new perceptual data, a blind person will really be able to

almost "see" the environment or 2D images [**91**]. Note that no depth information whatsoever is provided by this system. Due to the simplicity of the coding algorithm, the software for the vOICe does not require very high processing power and is already available as a JAVA applet for PCs or mobile phones. The equipment required for handling the vOICe consists in a glass-mounted camera, headphones, and a computer to perform the image-to-sound conversion, as shown in **Figure 2-35**.

Figure 2-35. A user wearing the vOICe system.

Capelle (1998)

Capelle *et al.* proposed the implementation of a crude model of the primary visual system in a hardware device [**92**]. The implemented device provides two resolution levels corresponding to an artificial central retina and an artificial peripheral retina, as in the real visual system. This prototype is based on a personal computer which is connected to a miniature head-fixed video camera and to headphones. Image processing achieves edge detection and graded resolution of a visual scene captured by the camera. Each picture element (pixel) of the resulting image is assigned a sinusoidal tone; weighted summation of these sinusoidal waves builds up a complex auditory signal which is transduced by the headphones. Note that the use of headphone, here and in the majority of works, obstructs the ears of the user, blocking out all environmental sounds.

At large, the auditory representation of an image is similar to that used in "TheVoice" [**78**] (described previously in **Figure 2-33**) just using distinct sinusoidal waves for each pixel in a column and each column being presented sequentially to the listener. Therefore, this

approach also suffers from issues such as complex sonification. The authors claim to have developed an on-line, real-time functioning prototype of a visual prosthesis. Also, they say that due to the ability to process real images in real-time the decisive learning phase is greatly enhanced, as well as sensory-motor interactions of users with their natural environment [**92**]. This conceptual model of sensory substitution has already undergone trials with success preliminary evaluation in a pattern recognition task during psychophysical experimentations, which demonstrated the usefulness of the present experimental prototype. We have created **Figure 2-36** (based on a similar figure presented in [**92**]) to depict the sensory substitution process proposed by Capelle.

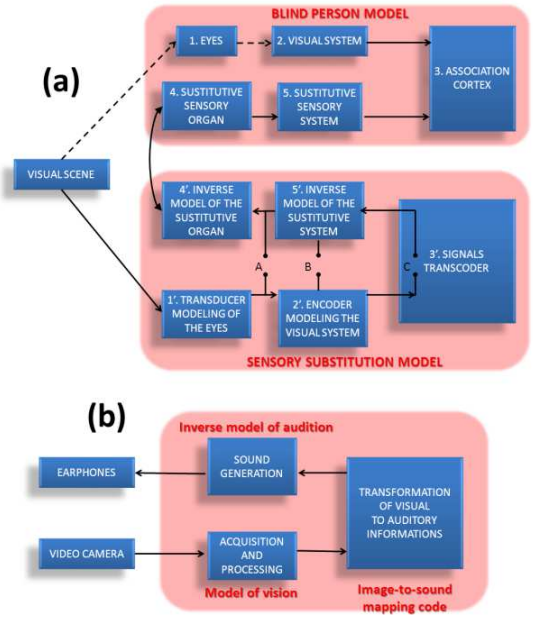


Figure 2-36. The sensory substitution process proposed by Capella et al. [**92**] in 1998.

The model depicted in this figure **Figure 2-36** was explained by the authors follow: "(a) Theoretical background. A model (1' and 2') of the visual system (1 and 2), processing visual information in a similar way, will theoretically produce a signal (3') comparable to the natural one reaching the association cortex (3). To allow this artificial signal (3') to reach the associative structures (3) through a substitutive sensory system (4 and 5), a model of this substitutive system (4' and 5') is reversed. By coupling (3') the model of vision (1' and 2') to the inverse model of the substitutive sensory system (4' and 5'), processing of visual information within the sensory substitution model (lower box) will produce a signal whose further processing by the natural substitutive sensory system (4 and 5) could supply the association

cortex (3) with the required signal. Coupling can take place at different levels on the pathway: at a theoretical level (3', link C), at an intermediate level (link B) if the knowledge of visual and substitutive systems are limited, or at a primary level (link A) if the connection is straightforward (case, e.g., of a tactile–vision substitution, or TVS, system [**76**], Philosophy and history of Sensory substitution). (b) Implementation of the sensory substitution model consisting of a model of vision connected to an inverse model of audition, using an appropriate transcription code". [**92**]

Gonzalez-Mora (1999)

Gonzalez-Mora *et al.* developed a prototype which incorporates video cameras and headphones mounted on a pair of spectacles, also spatialization of sound in the three dimensional space [**93**] as described in 3D sound and stereo vision. This device captures the form and the volume of the space in front of the blind person and sends this information, in the form of a sounds map, to the blind person through headphones in real time. The sound, therefore, is perceived as coming from somewhere in front of the user by means of head related transfer functions (HRTFs). The effect produced is comparable to perceiving the environment as if the objects were covered with small sound sources which are continuously and simultaneously emitting signals. The first device they achieved was capable [**93**]of producing a virtual acoustic space of 17x9x8 gray level pixels covering a distance of up to 4.5 meters. An aerial sketch of the system behavior for a particular example can be seen in **Figure 2-37**.

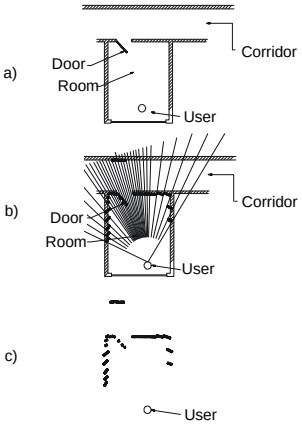


Figure 2-37. (taken from [93]) An aerial sketch of the system behavior.

In the part (a) of **Figure 2-37** the example of a simple environment is shown: a room with a half open door which leads to a corridor. A user is standing at the opposite side of the door (looking at it) with a window in his backside. In the sketch presented as (b) in **Figure 2-37**,

the result of dividing the field of view into 32 "stereopixels" is shown. This number denotes the lateral resolution (32°) of spatialization of this system through the horizontal axis. In other words, users can perceive sounds coming from 32 different positions from left to right. The final description of the environment is attained after calculation of the average depth (or distance) of each stereopixel. This description is then virtually converted into sound sources, located at every stereopixel distance. Therefore, a perception as shown in **Figure 2-37** (c) is produced. The authors claimed that out of this perception the user can identify the major components of the nearby space (the room itself, the half open door, the corridor, etc.).

In **Figure 2-37** the example has been constrained to 2-dimension for ease of representation. However, the prototype applies quite the same method to a third vertical axis. Thus, a 3-dimestional audio description of the images is given to the user at a rate of 10 images per second. Notice that this rate is fairly better than that of the vOICe, in which an image takes a second to be described. Though in this approach spatial information rather than visual information is provided, real time navigation is certainly feasible. The authors claimed that this new form of global and simultaneous perception of three-dimensional space via hearing, as opposed to vision, will improve the user's immediate knowledge of his/her interaction with the environment, giving the person more independence of orientation and mobility. Overall, this work opened the way for an interesting line of research in the field of the sensory rehabilitation, with immediate applications in the psychomotor development of children with congenital blindness.

Soundview (2003)

Soundview represents a single point of color as sound activated by the haptic exploration of an image on a tablet device [**94**]. Importantly though, Soundview does not use any haptic feedback (vibration, temperature etc.), but only the kinesthetic ability of the user to move through the image. To achieve color sonification, in Soundview the HSB (Hue, Saturation, Brightness) color space is represented through applying various filtering techniques to white noise (a combination of multiple frequencies) in an attempt to retain structural parallels with perceptual color space (**Figure 2-38**). White noise is filtered through a low pass filter, with brightness and velocity dependent on cutoff frequency[6] (Bright colors retain most frequencies, dark retain only lower frequencies). The result is filtered through a bank of 12 parallel reson filters[7] [**94**] (depending on hue and saturation) at octaves apart, a "Shepard filter" or "Shepard illusion" [**94**]. This allows a perceptually smooth transition between the highest (Hue = 0.99) and lowest (Hue = 0.01) frequencies in order to create a cyclic representation. The resonance frequencies, widths, and gains are dependent on the color and slide velocity [**94**].

[6] is a boundary in a system's frequency response at which energy flowing through the system begins to be reduced (attenuated or reflected) rather than passing through.

[7] a general-purpose filter that exhibits a single peak. The frequency response of a RESON filter is nowhere near as flat as a Butterworth filter frequency response. Nevertheless, all typical equalizer implementations typically implement classic reson filters.

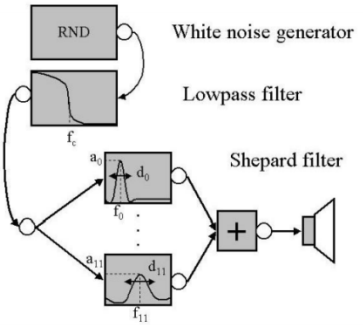


Figure 2-38. (taken from [**94**]) Filter graph to generate color dependent sounds. White noise is filtered through a low pass filter, with brightness and velocity dependent cutoff frequency f_c. The result is filtered through a bank of 12 parallel reson filters at octaves apart, a "Shepard filter". The resonance frequencies, widths, and gains are dependent on the color and slide velocity. [**94**]

An important contribution of this work is the design and implementation of an applet to tune various parameters that define the color to sound mapping interactively. The Applet can be run in Java 2 enabled for web browsers (**Figure 2-39**). At date, this SSD still lacks user studies to assess the usability of the system and therefore it holds unanswered all the open questions related to SSDs: Can people detect basic shapes with the system? How much detail can be perceived in this manner? How does performance on recognition tasks depend on training? Is there significant difference between blind and sighted people? These are examples of the questions that remain unanswered in this work. In fact, this system was never meant to be a mobility aid as commented by the authors who attempted the representation of only static images. It is of importance here to notice that like Bach-y-Rita and many others ([**71**], [**14**], [**5**], [**70**]), the authors claim that their system elicits synesthetic perception of an image through sound and touch. This belief rests on the assumption of theories related to brain plasticity and cross-modal transfer as discussed in Philosophy and history of Sensory substitution and Cross-Modal Transfer and Brain Plasticity sections.

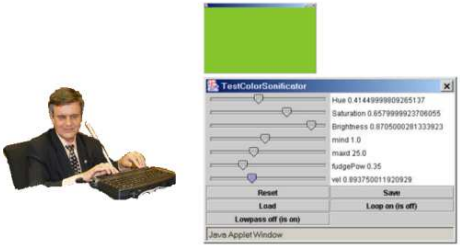


Figure 2-39. (taken from [94]) Soundview interface for interactive color-sound mapping.

TheVIBE (2008)

TheVIBE [**95**] is a visuo-auditory substitution system invented by S. Hanneton and currently developed in collaboration with Gipsa-Lab, France. The experimental device converts the video stream from a video camera into an auditory stream delivered to the user via headphones. Note here again the persisting issue (common among SSDs) of blocking out environmental sounds to the user by covering his (her) ears. The sound generated by *TheVibe* is a weighted summation of sinusoidal sounds produced by virtual "sources", corresponding each to a "receptive field" in the image [**95**]. More precisely, sets of pixels (white crosses, **Figure 2-40**) grouped in receptive fields (circled part, **Figure 2-40**) are distributed uniformly on the video-camera picture (right, **Figure 2-40**). Each receptive field drives the loudness of a particular sound source, which frequency and binaural panning are determined respectively by the vertical and horizontal position of the receptive field's center (white squares, **Figure 2-40**). One of the major originality of *TheVIBE* with respect to the other devices is that the video-to-sound mapping is entirely configurable.

Figure 2-40. (taken from [95]) The VIBE sonification model

The authors have tested *TheVIBE* on the mobility performance of twenty blindfolded subjects following a track in an ecological environment (in a maze on an indoor car park). Significant results were obtained with one hour of training only [**95**]. These results however were

more interesting from the designer's point of view than from the end user's one. Indeed, the authors recognize the lack of comparisons to assess how better the users perform when using *the VIBE*, and otherwise. Thus, assessment of the practical interest of this device for mobility requires indeed further investigations.

Kromophone (2009)

The Kromophone [**96**] is a program developed at Gordon College by Zach Capalbo to provide color information through an auditory stimulus. The SSD takes the input from a webcam and chooses the center pixel or an averaged area around the center pixel, and maps its color into several superimposed sounds. Each color is a sum of the focal sounds with the intensity (i.e. loudness) of each component determined by its contribution to the final color. So an orange color would be a mixture of red and yellow sounds, each of moderate loudness. In general, the intensity of the red is represented as a high pitch noise in the right ear, the green as a middle pitch noise in both ears, and the blue as a low pitch noise in the left ear [**96**]. The intensity of the white is then separated into distinct high pitch sound. The authors claim that the resulting sounds allow the user to distinguish colors.

Capalbo and Glenney conducted experiments with the kromophone [**96**] with both blind and blind-folded sighted subjects. They were able to identify fruit in a normally lit environment with only a few minutes to a few hours of training. Blind folded subjects were able to navigate the campuses sidewalks by distinguishing between the grey concrete and the green grass. Also, the authors suggested that the Kromophone outperforms the vOICe in search, discrimination and navigation tasks for blindfolded sighted participants. Luminance localization in a dark room had equal response times between the Kromophone and vOICe, however raising the ambient lighting conditions resulted in a sharp drop-off in performance for vOICe users only [**96**].

Michał Bujacz (2010)

The authors [**69**] mixed image processing methods with audio representation, to developed an algorithm for sonification of 3D scenes in an SSD as an aid for visually impaired individuals (**Figure 2-41**). The proposed algorithm includes the use of segmented 3D scene images, personalized spatial audio and musical sound codes. They used head related transfer functions (HRTFs) that were individually measured for all trial subjects. Thus, virtual sound sources originating from various scene elements were generated. By and large, the authors proposed to assign sound parameters to segmented object parameters, so as to distinguish the latter. The algorithm was implemented and tested in a SSD prototype with both sighted and visually impaired subjects [**69**].

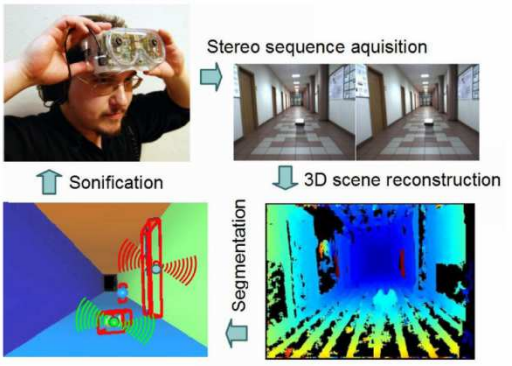


Figure 2-41. (taken from [69]) Scene sonification based on segmentation of an image stereo pair.

They use a depth-based algorithm to segment objects within a range map of the scene. This map arises out of a pair of stereoscopic images [**69**]. Once relevant objects have been segmented, the sonification algorithm attaches virtual sound sources to them (**Figure 2-41**), and presents them in such a way, that each is perceivable either as a separate auditory stream during training or as part of a familiar schema when scanning speeds were increased. Accordingly, the authors claim to give the visually impaired users a virtual sound environment that they have to scan in depth (**Figure 2-42**), so at to go through it safely. The depth-scanning proposed by the authors consists in a virtual scanning plane that moves through the scene (starting from the user and forwards) and releases sound sources as it intersects various scene segmented elements. This concept is shown in

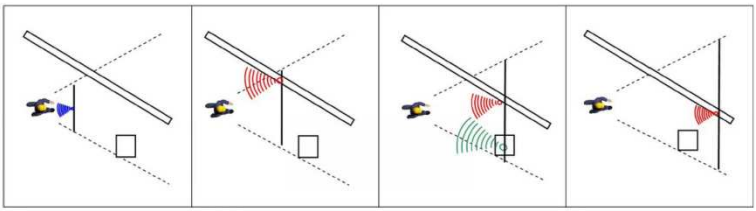


Figure 2-42. (taken from [69]) Scanning-plane progressing forwards in time (from left to right). [**69**]

Notice that in **Figure 2-42** the scanning-plane increasingly moves forwards as its size augments hand in hand with the perspective of the camera. Initially, no object is sonified since the plane intersections are non-existent. However, there is a sound in the middle of the

auditory field representing just the plane itself as it progresses forward. Later, sound sources start emitting as the plane intersects the associated objects. Distant objects are encoded with quieter sounds and larger objects are assigned longer sounds, encoding object's horizontal size with proportional duration.

EyeMusic (2012)

The EyeMusic [**33**] is a tool that provides visual information through a musical auditory experience. This SSD sonifies a 24x40 pixel colored image over 2 seconds [**33**]. Colors are first segregated into one of six categories (red, green, blue, yellow, white, black) and each category is encoded through timbre (e.g. White = piano, Blue = marimba). The color's vertical location is denoted through the pitch / note of the instrument, using notes across 8 octaves of a pentatonic scale (higher pitches reflect higher spatial positions) while the luminance of each color affects the loudness of the note (bright is loud and dark is quiet). Finally, each column is sequentially presented over time to complete the x axis. An illustration of the hardware and the mode of use of EyeMusic are shown in **Figure 2-43**.

Figure 2-43. (taken from [33]) Left: An illustration of the EyeMusic sensory–substitution device (SSD), showing a user with a camera mounted on his glasses, and using bone-conductance headphones, hearing musical notes that create a mental image of the visual scene in front of him. He is reaching for the red apple in a pile of green ones. Top right: close-up of the glasses–mounted camera and headphones; bottom right: hand-held camera pointed at the object of interest.

Recent experiments with the EyeMusic by Levy-Tzekdek *et al.* [**97**] examined the shared spatial representation of vision and EyeMusic soundscapes. Participants were required to indicate the location of an abstract target (either visually shown or heard using the EyeMusic) using a joystick. They found that altering the visual sensor-motor feedback (in this case, skewing the joystick's representation on screen by 30 degrees) influenced not only the visual trial movements but also the EyeMusic trials (where no feedback was given). The former

experiment indicates that soundscape feedback utilizing timing, timbre, loudness and pitch can inform a spatial representation for movement, while the latter experiment indicates a shared spatial representation [**97**].

Kai Wun (2012)

The authors in [**98**] presented a wearable stereo vision system for visually impaired. More precisely, this SDD is a glasses-like device intended to assist the blind to avoid obstacles in navigation. This device is composed of an eyeglasses and a power efficient embedded processing device. On the one hand, the eye glasses has embedded two miniature cameras for the stereo-imaging. And, on the other hand, a FPGA is used to synchronize and combine the stereo-images so as to get depth information [**98**]. The novelty of this work lies on series of parallel programming techniques, the device not only achieves a real-time stereo matching but, in addition, the video captured is streamed to a mobile device over the 3G network. This latter with the aim of enabling healthy sighted individual to remotely provide logistical guidance to the blind in real time. Therefore, the functionality of the device to avoid obstacles may be reinforced. Unfortunately, the authors fail to provide crucial information such as: how to detect obstacles out of depth maps and how to alert the blind in order not to bump into them.

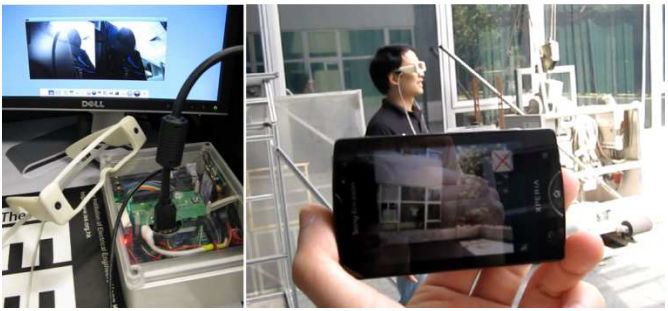


Figure 2-44. (taken from [**98**]) The setup of our wearable stereovision system (left). A real-time video streaming from a travel aid to a mobile phone (right).

AUGMENTATION OF REAL OBJECTS WITH VIRTUAL SOUNDS: ***The Shelfscanning (2009)***

The Shelfscanning [**99**] project is intended to empower visually impaired individuals to shop at their own convenience using computer vision methods such as object recognition, sign reading and text to speech codification. More precisely, the authors [**99**] advance in grocery detection using an object detection algorithm (**Figure 2-45**). Thus, ShelfScanner allows a visually impaired individual to shop at a grocery store without additional human assistance. To do so, online detection of items on a shopping list in video streams is performed. The

inputs out of which ShelfScanner detects an object, with the purpose of notifying the user, can be summarized as follow:

- Images of items from the user-supplied shopping list. These images come from the in vitro subset of the [**99**] GroZi dataset. The in vitro images are taken from the Web, and the images usually show ideal specimens of each product, taken under ideal conditions. The GroZi [**99**] dataset supplies between 2 and 14 in vitro images per item, with an average of 5.6 images per item.

- A video stream taken from the user's camera as she sweeps its FOV along shelves. The GroZi-120 video has resolution 720x480. Note in **Figure 2-45** that ShelfScanner relies on the assumption that the user holds a camera and sweeps the camera's field of view (FOV) across grocery shelves. Thus, a mosaic of assembled images is created. This is an important constrain for visually impaired individuals that the authors would like to alleviate in the future with a more general motion model.

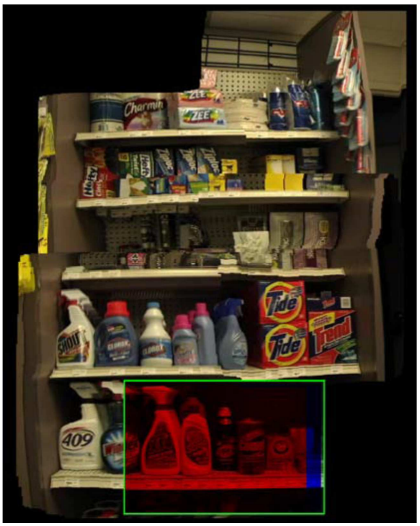


Figure 2-45. Grocery's identification using the Shelfscanner. [99]

The algorithm used in this project is based on offline supervised training [**100**] [**99**]. As for the online recognition of groceries the system: creates the mosaic representing the shelf, by assembling images from the camera. Then, points of interest (key points) are extracted and described using SURF descriptors [**100**]. Afterwards, a set of probability estimations are carried out, so as to decide whether a key point belongs to any of the sought targets (groceries in the dataset). The location of identified key points reflects the position of its associated

grocery. Finally, the visually impaired user is notified about the presence of groceries in his shopping list by natural speech.

NAVIG[8] (Navigation assisted by Artificial Vision and Global Navigation Satellite System).

The NAVIG [**101**] is an assistive device that can be roughly grouped into the electronic travel aids (ETA) family [**101**]. Its aim is to complement conventional mobility aids (i.e. white cane and guide dog) [**102**]. More specifically, NAVIG was designed to help visually impaired individuals cope with challenges at two levels: sensing their immediate environment (micro-navigation [**103**]) and reaching remote destinations (macro-navigation) [**102**]. On the one hand and quite like our See ColOr project, NAVIG uses Artificial Vision to detect objects which become reachable through 3D sound guidance. On the other hand, NAVIG also uses landmarks detection to refine a GPS-based user-positioning. This latter feature is a useful idea of data fusion that aims basically at matching GPS positioning, with current commercial GIS (Geographical Information System) information [**102**]. The end result is arguably a much reliable location which is compatible with assisted navigation for the Blind [**102**].

NAVIG's final prototype is a head-mounted stereo-camera system that operates on a laptop [**101**]. Although this prototype has been tested with relative success, to the best of our knowledge, more systematic and reproducible experiments are needed in order to draw better conclusions of this work. Despite the many similarities between this approach and ours (including the hardware display), we want to highlight what are perhaps their main distinctions:

- NAVIG cannot be classified as a Sensory Substitution Device, since no visual cue mapping is intended. NAVIG users are guided by mental imaginary or prior memories, yet the device does not promote visual-like experiences through the auditory pathway, as it is the case in synesthesia and SSDs.
- NAVIG is a device more concerned with outside exploration [**103**], which remains challenging in See ColOr.

[8] http://navig.irit.fr/

Figure 2-46. Navig project being tested: In outside navigation (left) and object grasping (right). Taken from [103]

Ribeiro (2012)

Rebeiro *et al.* [**8**] have 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. Therefore, they exploit the innate human capacity for 3D sound source localization and source separation, to orient the blind with respect to the objects nearby. A key aspect of this SSD concept (evaluated with a head-mounted device, **Figure 2-47**) is that unlike previous approaches, they use computer vision methods to identify high-level features of interest in an RGB-D stream. Thus, complex soundscapes aimed at encoding pixel-based representation of an entire image are avoided. The authors claim that since both visual and auditory senses are inherently spatial, their technique naturally maps between these two modalities, creating intuitive representations.

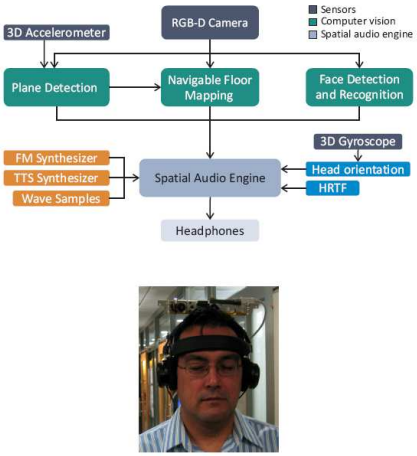


Figure 2-47. (taken from [8]) System block diagram (top) and device prototype (bottom). The synthesizers and wave samples are used to achieve the 3D spatialization of audio with respect of the user's head that is tracked with a Gyroscope.

This device is focused on the recognition of faces and planes only. On the one hand, plane detection is used to identify the floor, and to provide an environmental decomposition into flat surfaces. The authors claim that planes are the dominant primitives in manmade environments, and can be conveniently used to identify walls and furniture. Thus, the underlying assumption of the authors is that given the decomposition of an environment into planes of known sizes and locations (by sound), people can infer which classes of objects they belong to: "For instance, the location of a table could be inferred from the description of a large horizontal plane. Likewise, a chair could be associated with a small pair of horizontal and vertical planes" [**8**]. Consequently, they use an acoustical method for rendering planes based on 2D polar floorplan [**8**] (**Figure 2-47**). A 3D accelerometer is used to estimate the gravity vector, and infer the location of the floor plane. Also to correctly spatialize the sound with respect to the user, a head tracking is implemented with the use of a 3D gyroscope (**Figure 2-47**). On the other hand, finally, they recognize and represent them by the 3D spatialized name of the corresponding person. The device uses a musical note fallback, if a face is detected but not recognized.

2.3.4 General Discussion

There is an immense amount of information extractable out of sounds. Therefore, we see throughout the literature, the widespread idea of representing visual information to the auditory modality by systematically converting properties of vision (usually luminance, vertical and horizontal positions) into auditory properties (e.g. pitch, amplitude, frequency). Before the decade of the 90s, all the attempts to use such idea in synthetic auditory displays focused on the representation of simple shapes or visual patterns. Conversely, early in the 90's, the The vOICe (1992) [**78**] attempted to reach further by using this idea to encode natural images (pixel-by-pixel) into so-called soundscapes. The end result was a complex one-second sound whose amplitude and frequency synthetize the position and luminance of all the image pixels. Given the limitations of the auditory bandwidth with respect to the visual pathway, the use of soundscape in the vOICe gave rise to serious concerns about practical aspects of the vOICe. In principle, the users need to be extensively trained to derive visual meaning from a single soundscape, let alone of the many associated with all the frames in a video.

In this view, the authors claim that the usability of vOICe, rather relies on brain plasticity [**34**]. This is to say that with long-term practice the blind will learn to interpret the auditory scenes naturally. Several authors ([**14**], [**71**], [**5**], [**34**].) give support for this approach showing that after extensive training in soundscape identification, functional magnetic resonance images (fMRI) reveal that blind individuals present little activity in their primary visual cortex (i.e. cross-modal plasticity [**5**]). In fact, they observe that blind people demonstrate a shift in activated brain areas towards more posterior areas (the areas that are involved in visual processing in sighted people). However, whether this activity corresponds to actual visual experiences remains largely unknown, as it is the subject of even philosophical debates [**38**] on consciousness. Arguing against this, for instance, a plausible conjecture is that after 50 hours of training in recognizing the soundscape representative of an image, one is more likely to simply develop an associative pattern between the sound and the image, rather than having the visual experience as such.

Back to the evolutionary shaping of SSDs (in the late 90s' literature), researchers note that besides conveying visual information; the ability of the sight to (literally) focus on the information that is of relevance at any given time, also needed to be somehow modeled. For instance, understanding the environment layout so as to derive orientation is a skill which involves the dual abilities of localization and selective attention [**90**]. Notice that early attempts of auditory displays such as the vOICe, partially lack these features, since it is only coherent with the azimuth plane. In this spirit, authors such as Gonzalez-Mora (1999) [**93**] developed new SSDs by leveraging the idea of spatialization of sound and stereo vision. In fact, these two ideas made a difference in the history of SSDs evolution. While the former exploited the innate human capacity for 3D sound source localization and source separation, the latter provided depth information to virtually place these sound sources. Thus, the users are removed from the problem of inferring spatial information out of a soundscape (the vOICe), as it comes naturally to their ears. The use of these spatialized sound and stereo

vision combined in SDDs spans to the present days as seen in the works of Michał Bujacz (2010) [**104**] and Ribeiro (2012) [**8**], among many other.

By and large, few works in the literature attempt at providing a substitute to color information therefore, an important visual perceptual property is likely to be missed. The visual information and surrounding environments are mainly conformed by colored entities, so that their intelligibility could be diminished as colors are left out (more detailed information about color substitution in SSDs is given in Sensorial color coding). Recently in the last few years, one can find works that cope with this drawback in SSDs, namely those of Kromophone (2009) [**96**] and EyeMusic (2012) [**33**]. Although fairly good, these attempts clearly suffer from an issue that is often likened to the tunneling vision effect in retinitis pigmentosa [**105**]. More precisely, the exploration of the image is focalized on the center, making the field of view so narrow that achieving the general aspects of the image is rather unlikely. Further, the only explored part is hardly understandable as it lacks for context. This problem dramatically affects the overall understanding of an image (as a whole), because peripheral vision or global perception is unattainable.

A notable exception would be the Soundview (2003) [**94**] device (2003), which encodes color into sound applying various filtering techniques to white noise (a combination of multiple frequencies) in an attempt to retain structural parallels with perceptual HSL color space. This work largely overcomes the tunneling phenomena by enabling haptic exploration of the image on a tablet device (using the fingertips). Thus, in theory, the entire image resolution is made accessible (like in the vOICe), though only as many points as fingers can be accessed simultaneously. Which, in turn, avoids reaching the limits of the audio bandwidth (unlike the vOICe). In Figure 2-23 this modality can be identified as tactile sensory relocation, which implies by no means a tactile feedback, but a kinesthetic understanding of spatial relations of the image within the tablet (i.e. the finger serves to orientate the subject with respect to the picture). The kinesthesia refers to the innate ability to detect bodily position, weight, or movement of the muscles, limbs, tendons, and joints [**106**].

Despite increased use of computer vision nowadays, surprisingly, its use in SSDs is rather vapid. In others words, as noted by Jouffrais *et al.* in [**107**]: "*Although artificial vision systems could potentially provide very useful input to assistive devices for blind people, such devices are rarely used outside of laboratory experiments*". This is curious since one would think that devices meant to substitute natural vision should be heavily based on artificial vision. In fact, after reading works such as Ribeiro (2012) [**8**], NAVIG [**101**] and The Shelfscanning (2009) [**99**], one is left with the thought that if there is already a significant amount of research on methods to enable computers to "see", these very methods might be targeted to the benefit of blind individuals. 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. For instance, NAVIG [**101**] is indeed a promising approach very akin to our concept of assistance and mobility. Yet, it still lacks testing and developing, and more importantly, in our view, it should somehow account for visual substitution beyond its orientation/guidance power. Also,

the work presented by Winlock *et al.* The Shelfscanning (2009) [**99**], though single-task oriented, is a good example of how machine learning and image processing can be oriented to aid the visually impaired in daily activities, such as grocery shopping. Similarly, the work presented by Ribeiro *et al.* Ribeiro (2012) [**8**] make efficient use of face detection methods to communicate identities to the blind via spatialized speech. These ideas could be extended to a broader gamut of objects and hence tasks, to increase in independence of the blind. In fact, there is another emergent project that will attempt this line of thinking: the "fifth sense project", sponsored by the Andrea Botticelli foundation and the M.I.T (Michigan Institute of technology) [**108**]. Yet no intellectual production (or literature) has been brought to light out of this developing idea. And similarly to NAVIG [**101**], it seems that this project won't integrate visual substitution either.

Vision embraces more than the sensing of isolated low-level visual features (luminance, contrast, texture, etc.). In fact, besides information acquisition, vision is rather a holistic phenomenon emerging out of the integration of all the visual information (shape, color, position, distance, texture, luminance etc.) [**16**]. Even further, 'normal' vision is itself constrained by top-down knowledge, so it would be unpractical to deny to this knowledge a role in visual sensory substitution [**14**]. Top-down knowledge produces the kind of information that sighted individuals achieve from their visual systems, typically without conscious effort [**108**]. For instance, the recognition of faces is an unconscious process that our brain performs automatically based on previous concepts. In other words, we do not have to re-apprehend the aspect of faces every time see one. This top-down knowledge, or unconscious knowledge, or accumulative information, could not be better emulated in sensory substitution by other than computer or artificial vision. Also, we maintain that speech (or alternatively, earcons[9]) is a valid method for object sonification in SDDs, as it is one of the most powerful (and the most used) form of auditory communication.

Those maintaining that SSDs elicit an acquired synesthesia (showing activity in the visual cortex [**14**], [**71**], [**5**]), may argue that speech-based descriptions trigger just visual imagery rather than vision via sensory substitution. In fact, Ward *et al.* [**14**] say that while a car horn evokes the image of a car, this is very different in nature from the information in a soundscape (produced by an algorithm that maps an image 'containing a car' into sound). The horn sound turns out to be general symbolic mapping mediated by the concept 'car', whereas the soundscape may convey specific information of the scene, car's type, perspective, location and so forth. However, others like O'Regan [**35**] argue otherwise: "The only difference is that whereas imagining finds its information in memory, seeing finds it in the environment. One could say that vision is a form of imagining, augmented by the real world."

On our side, we advocate the use of natural speech to label objects, against soundscapes, as it prevents users from spending 70 hours of training (or more) [**14**] in recognizing an object. The end result should be a SSD absolutely learnable, intuitive and extremely practical.

[9] a brief, distinctive sound used to represent a specific event or convey other information.

Here practical refers not only to the ease of use, but to the efficiency in handling user's prior knowledge to avoid tough learning processes. In congenitally blind their concept of the objects whatever it is (e.g. tactile) is exploited, and in people who became blind their visual imagery is tapped. Further, SDDs may offer better insight into the scene by spatializing the speech to represent left/right, modifying the pitch to represent top/bottom, and adjusting the volume to represent the depth of the objects. Thus, a scene presenting various objects would convey a great deal of the information expected into a soundscape, though preserving simplicity and intuitiveness. We found these ideas often abandoned in the literature.

Otherwise, there are certain devices that have a family resemblance to SSDs but would not meet the more restricted definition of sensory substitution [**14**]. Here, however, we would like to do a brief description of them since those devices are meant to aid the visually impaired in navigation and exploration. More importantly, they are intended to enable a degree of 'normal' functioning for the blind (the main topic of this work). White canes and dogs are established aids for mobility. They are both used by visually impaired persons for near space perception tasks, such as orientation and obstacle avoidance. Nevertheless, a dog has a significant cost and can assist mobility ten years, on average. According to Hersh [**109**], mobility aids can be classified by the nature of the assistance provided (mobility or orientation), the space which will be explored and the complexity of the technology [**109**]. The main classes are:

- traditional low-tech aids for near space exploration;
- electronic travel aids of medium-tech for near space exploration;
- high-tech electronic orientation aids for large space exploration;
- mobility aids for near/far space navigation.

One of the most representative examples that efficiently combines all the above categories is the work presented by Pissaloux et al. *SEES* (Smart Environmental Explorer Stick) [**110**], [**111**]: an enhanced white cane which assists the navigation of the visually impaired. This active multi-sensor context-awareness concept integrates three main devices: iSEE (global remote server), SEE-phone [**112**] (an embedded local server) and SEE-stick (smart stick); which complements each other. This novel idea is yet to be fully implemented and tested, though preliminary studies show its promising potential.

Two essential constituents of this classification are exploration and navigation. Exploring a particular place means discovering the main components or looking for something in it. A navigation task for a visually impaired person involves making a decision on which course to follow between a starting point and a destination point. Note that obstacles should be avoided for both tasks. An example of the first class is the white cane, while in the second we find among others, several variants of a cane having laser or ultrasound sensors that provide the user with distance measurements translated into tactile or auditory data. Examples include: LaserCane [**113**], GuideCane [**114**], UltraCane [**115**], and Télétact [**116**].

High-tech electronic orientation aids for large space exploration assist visually impaired individuals in tasks such as self-localization and space perception by verbal description. The Global Positioning System (GPS) is a fundamental ingredient of the prototype aids belonging to this class. However, the GPS provides an inaccurate localization (about 10–15 m) and the signal is absent indoor and underground. The Mobic travel Aid, [**117**] [**118**], the Sextant System [**119**], and Loomis' Personal Guidance Systems [**120**], are examples of this class of assistance. A more recent system is the Braille Note GPS [**121**]. In the future it will be possible to obtain better localization, as the forth- coming Galileo Global Positioning System will provide two levels of precision: 1m for commercial applications and 5m for public use.

For the last class of mobility aids, the GUIDO Robotic Smart Walker is a representative example that serves as support and navigation aid [**122**]. This prototype builds maps of the close environment with depth captured by a laser. The assistant robot can calculate the path from one point to another and can also avoid obstacles. Another example is represented by Talking Signs [**123**], which is an information system consisting of infrared transmitters conveying speech messages to small receivers carried by blind travelers. A user can get to the destination by walking in the direction from which the message is received. Drishti is an integrated indoor/outdoor blind navigation system [**124**]. Indoors, it uses a precise position measurement system, a wireless connection, a wearable computer and a vocal communication interface to guide blind users. Outdoors, it uses a differential GPS to keep the user as close as possible to the central line of sidewalks. Moreover, Drishti uses a Geographical Information Systems (GIS), in order to provide the user with a speech description of the close environment.

Moreover, a large number of tools have been created so far to help bind people in tasks others than navigation/exploration, such as the perception of texts [**125**], pictures or chart data [**126**] [**127**] [**128**]. In fact, audio has been used extensively to present spatial information to blind and visually impaired users [**129**]. There are many examples of using audio to represent the shape of graphs and charts, which are traditionally heavily visual methods of presenting information.

Brown and Brewster [**130**] describe how pan and pitch can be used to represent x, y values, respectively, when displaying an audio line graph. They were able to demonstrate that users can accurately draw shapes of the graphs they hear. Alty and Rigas [**131**]describe their AUDIOGRAPH system that can be used to display simple diagrams to users. They vary the pitch of two tones of different timbre to represent an x, y position in a 2D space. They demonstrate that it is possible to display simple shapes to a user through varying the pitch of audio tones.

Zhao *et al.* [**132**] use audio to display geographic information through active exploration, using a tablet (or keyboard)-based map system. They divide a map hierarchically into regions varying timbre, pitch, and pan to display a value for a region, along with the altitude and azimuth information. Percussive sounds are played to alert users when they move between

regions. Kamel *et al.* [**133**]combine audio feedback with a tablet for input to display simple graphical shapes. Users explore the shapes using a tablet, with audio cues alerting them when they enter different areas of the diagram. Shapes are represented by nonspeech sound sources moving through space (using 3D audio techniques to move the source horizontally and vertically). Users can track the movement of the sound to recreate the shape. Changing the pitch of the sound supports the user for vertical movements, as it has been shown that users have difficulty placing sources in the vertical dimension when they are presented using standard 3D audio-rendering techniques.

To conclude this section we would like to compare our approach See ColOr (to be presented next) with the most relevant SSDs cited so far in this work. We generated a table (Table **2-1**) that encompasses the main features expected to be preserved as images are converted into sounds, and that at the best of our knowledge, constitute the makings of an efficient SSD. These features range from low-level such as, spatial perception within the azimuth plane (x axis), awareness of elevation (y-axis) and depth (z-axis). Although, we also track in this table high-level visual features often liken with unconscious workings of the brain, through which we derive sense from the visual world (i.e. automatic recognition of known objects and faces, awareness of obstacle nearby, and perspective acquisition). Otherwise, this table (Table **2-1**) further covers a number of technical sides relevant to sonification that ultimately yield agreeable user experiences. In this work we are concerned to know whether a SSD permits the exploration of more than one visual point simultaneously. Also, the use of soundscapes significantly matters into the scope of this work, as the reactivity of SSDs (real time) heavily depends on this fact. Last but not least, two aspects will be considered (Table **2-1**): firstly, whether the sound is conveyed to the user at expense of blocking out his ears, or by means of bone-conduction techniques that prevent environmental sound obstruction. Secondly, we would like to assess whether or not current SSDs make use of blind user's interfaces to allow better interaction with the system.

We will observe in this table (Table **2-1**) that See ColOr meets sufficient conditions to support its usability. Particularly, the sonification of elevation is just partially achieved in See ColOr: while sound is not treated to reflect this feature, the tactile reference frame offered by our system (Haptic-based Interfacing) will cope with this drawback in great extent. Furthermore, alternative solutions to this drawback in See ColOr will be discus in section Improving time in experiments. Soundscape usage is another missing feature in See ColOr, yet as discussed earlier, this is not necessarily a disadvantage. We will argue in this thesis (But is See ColOr functional so far?) that more efficient methods may be adapted for object sonification, on the grounds that soundscapes are extremely tough to understand and dramatically reduce reactivity of the system. In addition, out of this comparative table (Table **2-1**) it will be clear that the perspective effect continues to be challenging for SSDs, as all of them fail to reproduce it. Note that this table (Table **2-1**) also reflects the fact of See ColOr being unique in combining both, low-level and high-level features into a SSD.

	X axis (azimuth)	Y axis (elevation)	Z axis (depth)	Multiple points	Soundscape	Color	Objects	Obstacles	Faces	Perspective	Real Time	craneal conduction	Interface
Raymond M. Fish													
The VoiCe													
Capelle *et al.*													
Gonzalez-Mora *et al.*													
Soundview													
TheVIBE													
Kromophone													
Shelfscanning													
Michał Bujacz *et al.*													
EyeMusic													
Kai Wun *et al.*													
Ribeiro *et al.*													
Fifth sense													
See ColOr													

Table 2-1. Table features that serve to compare efficiency between SSDs. Red stands for missing features, whereas green means that the feature is met.

3 SEEING COLORS WITH AN ORCHESTRA

Throughout this chapter the system for auditory visual substitution proposed in this thesis will be presented in details. The 3.1 Overview section fully depicts the workings of this aid for the visually impaired, whereas the 3.2 Evolution section relates to the technical transitions the system has gone through. Thereafter, the following sub sections provide a precise explanation on sonification methods (see 3.3 Sonification), tactile interaction (see 3.5 Haptic-based Interfacing), optimal handling of optical sensors (see 3.4 Efficient registration of imaging sensors), and computer vision strategies (see 3.5 Computer-vision-based visual substitution). All this together makes up a holistic approach underlying the functioning of the system put forward in this document. Though this thesis affords a whole chapter for selected experiments, testing with users and experimental checkups may be encountered too across this chapter. Overall, this chapter 3 describes in details the mathematical approaches, experimental setups, frameworks, hypothesis, computational techniques and practical methodologies used in this thesis.

3.1 Overview

As Pissaloux *et al.* say: "visually impaired people need to improve both orientation and mobility capabilities in order to be able to walk autonomously and safely" [**110**]. This being said, we must recall that these tasks take more than simple obstacle detection and avoidance. They require perception and understanding of the whole nearby environment. As a result, mobility, for example, encompasses three main tasks. The first, is to understand the near space global geometry; the second, is simply to walk and to avoid obstacles; and the third is to walk with a specific goal to reach, for instance looking for a specific door, a shop entrance, etc. Hence, researchers in vision and visual (cognitive) systems such as Pissaloux, advocate the design of holistic technological/computational approaches to the concept of mobility [**134**].

Others, like Jouffrais [**107**], also argue that bio-inspired computational models must be target to restore essential visuomotor behaviors that allow the blind to robustly, precisely and rapidly locating visual cues in the environment. In this view, we put forward **See ColOr** (whose name stands for: Seeing Colors with an Orchestra) a mobility aid for visually impaired people that use the auditory channel to represent portions of captured images in real time. A distinctive feature of the See ColOr interface is the simultaneous coding of color and depth. Also, See ColOr now uses computer vision methods for processing more complex visual features eventually sonified as virtual objects/obstacles. Four main modules were developed, in order to replicate a number of mechanisms present in the human visual systems. These modules are:

- the local perception module;
- the global perception module;
- the alerting system;
- the recognition module.

The **local module** provides the user with the auditory representation of a row containing 25 points of the central part of captured image. These points are coded into left-right spatialised musical instrument sounds (Sensorial color coding), in order to represent and emphasize the color and location of visual entities in the environment. The key idea is to represent a pixel as a sound source located at a particular azimuth angle (How does See ColOr sound like?). Moreover, each emitted sound is assigned to a musical instrument (From colors to instruments sounds in See ColOr), and to a sound duration depending on the color and the depth of the pixel, respectively (The sound of local and global modules). The local module allows the user to explore a captured video frame indeed. However, since the perception of the user is focused only on a small portion of the captured scene (single point), the tunneling vision phenomenon becomes apparent (The sound of local and global modules). Having access to more than one point would bring many advantages in terms of exploration and understanding of the image. For instance, it would be possible to compare several distant points (regarding color and depth) if pointed with the fingers on a touchpad displaying the environment picture. 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 exploring points beyond the image center.

In the **global module** the image is made accessible on a tactile-tablet interface (Haptic-based Interfacing) that makes it possible for the user to compare different points and explore the scene in a broader context (The sound of local and global modules). 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. The spatial position of this pixel (in the azimuth plane) is mapped to the hearing by directional virtual sound sources (The sound of local and global modules). In theory, the entire image resolution (460x640 using the Kinect camera) is made accessible so as to make the most of the camera information. However, only as many points as fingers can be accessed simultaneously not to reach the limits of the audio bandwidth

(Optimal interaction). 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. This is mostly the case because (s)he can access various points simultaneously (several fingers), slide a finger so as to scan areas and so forth (Building a scene in someone's mind).

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. The purpose of **alerting system** (Obstacles detection) is to warn the user whenever a hazardous situation arises from obstacles lying ahead. 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. 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 (Obstacles detection). It is worth noticing that the alerting system is an autonomous algorithm (Computer-vision-based visual substitution) 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.

See ColOr also uses computer vision techniques (Computer-vision-based visual substitution) to process higher visual features of the images in order to produce acoustic virtual objects. Actually, we recognize and then sonify objects that do not intrinsically produce sound, with the purpose of revealing their nature and location to the user. The **recognition module** is a detecting-and-tracking hybrid method for learning the appearance of natural objects in unconstrained video streams (Object recognition). Firstly, there is a training phase aimed at learning the distinct appearance of an object of interest (scale, rotation, perspective, etc.) (Learning). This is an off-line process carried out by sighted people. Then, a visually impaired individual may be informed about the presence of learned objects in real time during exploration, if any (Running). Overall, this module allows the blind noticing serendipitously discoveries; seeking a specific target; and avoiding obstacles as well.

In **Figure 3-1** we have shown an illustrative drawing of a user endeavoring to navigate/explore an unknown environment with the help of See ColOr. By default See ColOr runs in **local module**, thus only upon request when tapping on the tablet, the **global module** enters activity. In **Figure 3-1** for instance, the user is tapping the right-up corner of the chair's back (on the tablet), as he explores the picture. Therefore, he is expected to deduce out of the emitted sound, the color and distance (depth) of that part of the chair. Note that the user hears the sound of a touched point on the tablet like it is coming from its peer in the real world e.g. from left to right or azimuth plane, but not in elevation. Moreover, while the user walks, the **alerting system,** running in parallel, will automatically announce (by means of an alarm) all potential obstacles on his way e.g. the garbage can. Eventually, the **recognition module** is meant to reveal known objects happen to be nearby. In **Figure 3-1**, it is

assumed that See ColOr previously learned to recognize a chair. Although the **recognition module** might learn a broader gamut of objects, including faces and people identities.

Figure 3-1. See ColOr's overall functioning.

3.2 Evolution

See ColOr is a project originally founded by Swiss Hasler Foundation and developed at the Computer Vision and Multimedia Laboratory (CVML) of the University of Geneva and the School of Engineering, Switzerland. Beginnings of this project date back to 2005, though it was put on hold in 2008 for over two years. When retaken in 2010, See ColOr barely had implemented its local module. Many years have passed and the shaping of a functional prototype in time could be summarized as in **Figure 3-2**. Early attempts of a tactile interface were achieved by means of a Talking Tactile Tablet or T3. This is a graphic tablet with a touch surface that uses swell paper to create 3D overlays and connects audio-files to parts of the overlays. The device is connected to a computer and has a tactile surface which produces touchable icons that provide audio feedback when they are pressed. Specifically, this tablet makes it possible to determine the coordinates of a contact point. This tablet had dimensions of 15"L x 12"W x 1.5"D, and a weight of 6.5 lbs (2.94 kg). More recently, in 2011, we made used of a superlight and much smaller Pen Tablet (Bamboo Wacom). When connected to a computer, this tablet bridges a user and an interface (picture) by means of the fingertip (which acts as mouse pointer). Nowadays, we use an iPad tablet (Apple) whose advantages are as follow: wireless connection, higher resolution of touch, unlimited multi-touch, omnidirectional adaptive rendering, and in the near future, it is expected to replace the laptop and carry out computational processes (see **Figure 3-2**).

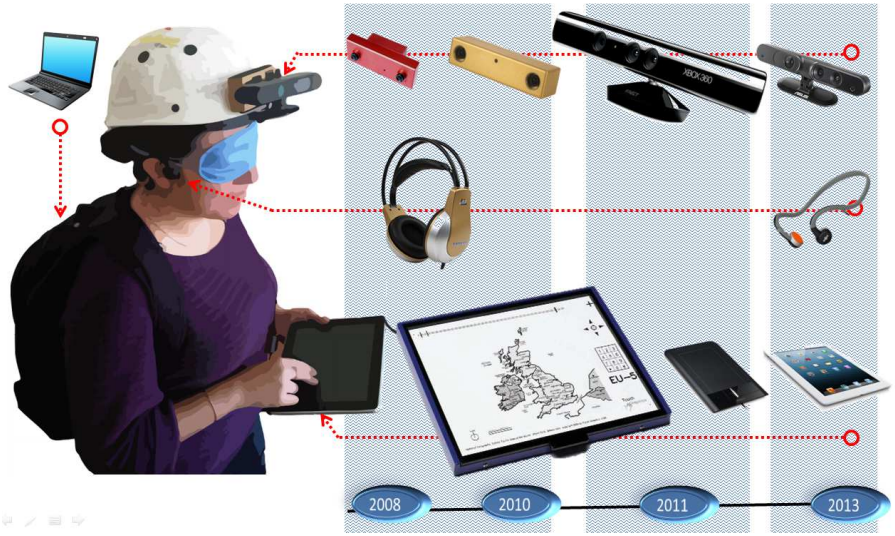


Figure 3-2. Evolution of See ColOr's prototype.

The use of earphones was a problem in See ColOr during many years (see **Figure 3-2**). In fact, this is an issue widespread in auditory-based substitution of vision. At large, people are reluctant to block out their ears even in exchange for assistance: "I already miss a sense and wearing headphones is like taking one more way", that is a commonplace comment among users. Indeed, natural audio cues of the environment relevant for self-orientation are likely to be missed in this way. See ColOr surpassed this difficulty, recently, with the addition of bone-conducted sound in the current prototype. This strategy turned out of broad acceptance among users, a key aspect to add to the functional features of See Color. Bone conduction is the conduction of sound to the inner ear through the bones of the skull using an ergonomic vibrating headset (bone-phones). This device converts electric signals into mechanical vibrations and then, sends transduced sound to the internal ear through the cranial bones. Since the outer ears are no longer involved in the process of hearing, they remain released and uncovered. Note that in **Figure 3-2**, the user wears the bone-phones right over the temples, rather than in/on the ears.

Over the years See ColOr has used various range cameras as optical sensors (**Figure 3-2**). In chronological order, the most representative ones are: all-digital stereo MEGA-DCS camera (Videre Design), The Bumblebee®2 stereo vision camera (Point Grey), the Microsoft 3D sensor Kinect (SDK), and lately, the ASUS Xtion PRO LIVE (SDK). The first two sensors are stereoscopic-based, programmable, quite small, long ranged and outdoor functional. Yet, they provide only sparse depth maps leading to high levels of depth uncertainty. Further, these

cameras require IEEE-1394 (FireWire) connection, which severely limits their use let alone of elevated prices. Recently, cost-efficient sensors using Time-of-flight technology, brought along compact depth maps and flexible programing frameworks (Kinect). See ColOr, as many research projects, made use of this sensor for long period. However, this camera suffers from both, oversize and power feeding need. This latter forced us in See ColOr to build an unsightly connection based on an 8 pounds rechargeable battery that needed to be carried by the user. Nevertheless, the later generation of this camera (ASUS Xtion Pro live 2012) was fashioned smaller, lighter and based on plug-and-play USB connection. Therefore, See ColOr currently uses this 3D sensor to capture both color images and depth maps of the environment. Importantly though, since the resolution of the internal color camera may not be sufficient for computer vision applications. In See ColOr the addition of an external high-resolution camera (webcam) as shown in **Figure 3-3**, is possible (see Efficient registration of range and color images).

Figure 3-3. Coupling of a 3D sensor (ASUS Xtion Pro Live) with an external camera (regular webcam) to increase color resolution for better performance in computer vision applications.

3.3 Sonification

3.3.1 Sensorial color coding

In the world, visual information and surrounding environments are mainly conformed by colored entities; therefore their intelligibility could be diminished as colors are left out of SSDs. In fact, color plays a very important role in our everyday lives. For instance, we can judge how healthy we, our crops and our food are, by means of the color. We made choices (decoration, furnishing, clothing etc.) being highly influenced by colors. One could say, therefore, that color is involved in almost every aspect of our lives. More specifically, there are three key aspects why the blind and visually impaired may benefit from the coding of color in SSDs. Firstly, colors provide clues to object identification (e.g. apples are red and bananas are yellow). Secondly, the accessibility to color information permits communication between the blind and sighted about the visual world, so they share concepts on similar basis. Finally, color information is very relevant for figure-ground segmentation ([**135**] [**136**] [**137**]).

In 2009, Torralba et al. [**137**] found that the importance of color information for visual scene/object recognition is greater for medium and low resolutions than it is for higher resolutions [**137**]. This is of much importance for SSDs, as most of them have a relatively low resolution being constrained by both technology limitations and the users' perceptual ability. In tactile devices the limit is set by the numbers of stimulators (hardware), whereas in auditory systems the limitation is given by the band-width of the ears that is typically low.

In this view, research on sensorial color coding for SSDs continues to gain significance nowadays. Recently, the research domain of color sonification has started to grow [**94**] [**138**] [**96**]. A number of authors defined sound/color mappings with respect to the HSL color system. HSL (Hue, Saturation, Luminosity) is a symmetric double cone symmetrical to lightness and darkness. HSL mimics the painter way of thinking with the use of a painter tablet for adjusting the purity of colors. The H variable represents hue from red to purple (red, orange, yellow, green, cyan, blue, purple), the second one is saturation, which represents the purity of the related color and the third variable represents luminosity.

The H, S, and L variables are defined between 0 and 1. Doel defined color/sound associations based on the HSL color system [**94**]. In this sonification model, sound depends on the color of the image at a particular location, as well as the speed of the pointer motion. Sound generation is achieved by subtractive synthesis. Specifically, the sound for grayscale colors is produced by filtering a white noise source with a low pass filter with a cutoff frequency that depends on the brightness. Color is added by a second filter, which is parameterized by hue and saturation.

Rossi *et al*. presented the "Col.diesis" project [**138**]. Here the basic idea is to associate colors to a melody played by an instrument. For a given color, darker colors are produced by lower pitch frequencies. Based on the statistics of more than 700 people, they produced a table, which summarizes how individuals associate colors to musical instruments. It turned out that the mapping is: yellow for vibraphone or flute; green for flute; orange for banjo or marimba; purple for cello or organ; blue for piano, trumpet or clarinet; red for guitar or electric guitar.

Capalbo and Glenney introduced the "KromoPhone", whose general aspects were already discussed in this work (Kromophone (2009)) [**96**]. In terms of color treatment, their prototype can be used either in RGB mode or HSL mode. Using HSL, hue is sonified by sinusoidal sound pitch, saturation is associated to sound panning and luminosity is related to sound volume. The authors stated that only those individuals with perfect pitch perform well. In RGB mode the mapping of colors to sounds are defined by pan, pitch and volume. For instance, the gray scale from black to white is panned to the center, with black being associated to the lowest pitch sound. Blue and yellow are mapped to the left, with blue being associated to lower pitch than yellow. Similarly, green and red are related to sounds listened to the right. Finally, the intensity of each color is mapped to the volume of the sound it produces.

In one of their experiments, Capalbo and Glenney illustrated that the use of color information in a recognition task outperformed the performance of "TheVoice" (The vOICe (1992)) [**96**]. Specifically, the purpose was to pick certain fruits and vegetables known to correlate with certain colors. One of the results was that none of the three subjects trained with "TheVoice" could identify any of the fruit, either by the shape contours or luminance, while with the "KromoPhone" individuals obtained excellent results.

Meers and Ward proposed the ENVS system which code colors by means of electro-tactile pulses [**139**] [**140**]. Note also that they consider color perception very important, as it can facilitate the recognition of significant objects which can serve as landmarks when navigating the environment. As stated by the authors, delivery of all colors by frequencies proved too complex for accurate interpretation. Consequently, in the ENVS prototype only an eight-entry lookup table was implemented for mapping eight colors.

It turns out that if we would like to use one of the sensorial color coding described above, we would come across several limitations. Specifically, we would like to use a system that reacts in real-time; thus, the sonification used in the Col.diesis project would be unsuitable, since the sonification of a single color last many seconds [**138**]. The KromoPhone color sonification is very reactive; however, because of the pan color coding, only a single pixel could be sonified. In other words, the spatial coding of more than a pixel would not be possible. As we would like to represent simultaneously a reasonable number of sonified pixels with also their corresponding spatial positions, we also dispose of this color sonification scheme.

A similar argument yields the same conclusion for Doel's system, which is also much more oriented toward the exploration of static pictures. The colors/electro-tactile mappings of the ENVS system present the advantage to represent ten pixels, simultaneously. However, in terms of quantity of information the tactile sensorial channel represents 0.1 Kbps, while audition is about 10 Kbps [**6**]. Thus, we prefer to represent colors by sounds transmitted to the auditory pathway. Finally, we would like also to represent color and depth, simultaneously. With a tactile interface it is unclear to us on how to achieve an efficient coding in real-time.

Sjöström and Rassmus-Gröhn [**141**] reported a computer-haptic interface using PHANToM – a joystick that moves in a 3D space or 'haptic scene'. Different colors are represented by different levels of resistance as the device is moved thereby creating a sense of texture. Moreover, Cappelletti, Ferri and Nicoletti [**142**] represented a single point in RGB color space on three fingertips (relating to red, green and blue dimensions) and by varying vibrotactile frequency (low, medium, high amplitude). So the presence of red would be felt as high vibration intensity on the 'red finger', yellow would be felt as high vibration intensity on the red and green fingers (because yellow is represented by 1, 1, 0 in RGB space), and so on.

Finally, the Chromo-Haptic Sensor-Tactor (CHST) device has four short-range fingertip-mounted color sensors as part of a glove [**143**]. The four finger-mounted sensors are tuned to

detect four dimensions each: R, G, B and luminance (allowing 1 to 4 different colors simultaneously). The information is relayed to four belt-mounted vibrotactile stimulators (T1-4) [**143**], varying in vibration and temporal modulation to convey color information. Unlike the initial attempts to transfer color information, this device is the first to use a color sensor for the external environment rather than a pc camera.

3.3.2 From colors to instruments sounds in See ColOr

The goal is to use the auditory pathway to convey color information as quickly as possible. The simplest method would consist to use human voice. Nevertheless, having a voice announcing the colors leads to various difficulties associated with interpreting common speech. The main problem is that we would like to communicate several pixel colors, simultaneously. Note that for a person it is almost impossible to follow at the same time a discussion with more than three individuals. In other words, processing speech unconsciously requires a lot of mental resources. It may be possible to understand the name of two colors, though saying a color takes about a second or even more if the palette has a considerable size, which is also too long for real-time purposes. Thus, the lengthy spelling of names might reduce the information transfer rate between the visually impaired individual and the system. Finally, it would be difficult to remember the name of hundreds of different hues.

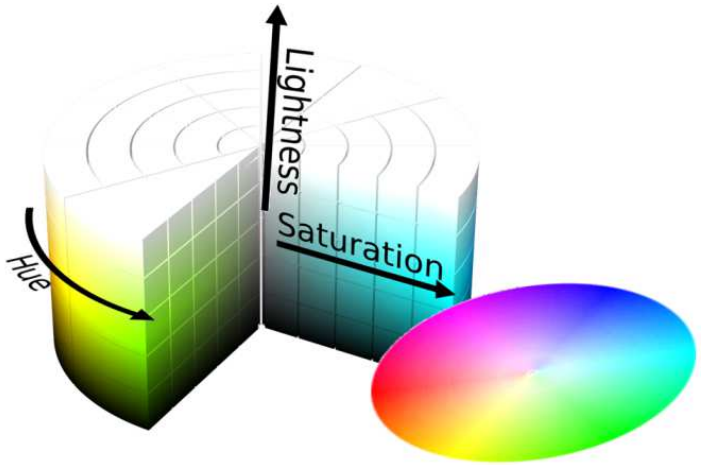


Figure 3-4. The HSL system used in See ColOr. In this figure the slice extracted from the middle of the cylinder is meant to show the Hue variable with neutral lightness (basic colors).

An intuitive approach entails the sonification of color system variables as seen in Sensorial color coding. The RGB (red, green, and blue) cube is an additive color model defined by mixing red, green and blue channels. For instance, we used the eight colors defined on the vertex of the RGB cube (red, green, blue, yellow, cyan, purple, black and white). In practice a pixel in the RGB cube was approximated with the colour corresponding to the nearest vertex. Our eight colors were played on two octaves: Do, Sol, Si, Re, Mi, Fa, La, Do. Note that each color is both associated with an instrument and a unique note. An important drawback of this model was that similar colors at the human perceptual level could result considerably further on the RGB cube and thus generated perceptually distant instrument sounds. Therefore, after preliminary experiments associating colors and instrument sounds we decided to discard the RGB model.

In this view, a number of authors defined sound/color mappings with respect to the HSL color system (see Sensorial color coding). The HSL color system also called HLS or HSI is very similar to HSV. Advantages of HSL are that it is symmetrical to lightness and darkness, which is not the case with HSV.

As shown by **Figure 3-4**, the HSL color system (Hue, Saturation and Luminosity) may be regarded as a cylinder extended from lightness down to darkness. Although the HSL color system is related to the laws of physics and not to the human perception system, it is much more intuitive than RGB. HSL mimics the painter way of thinking with the use of a painter tablet for adjusting the purity of colors. Other color systems such as Lab or Luv are strongly related to the human perception system, but the difficulty lies in the determination of an intuitive matching between color variables and sounds.

In the HSL color system the H variable represents hue from red to purple (see **Figure 3-5**), the second one is saturation which represents the purity of the related color and the third variable represents luminosity. Hue varies between 0 and 360 degrees (see **Figure 3-5**), while S, and L are defined between 0 and 1. We represent the hue variable by instrument timbre, because it is well accepted in the musical community that the color of music lives in the timbre of performing instruments. Moreover, learning to associate instrument timbres to colors is easier than learning to associate for instance pitch frequencies. The saturation variable S representing the degree of purity of hue is rendered by sound pitch, while luminosity is represented by double bass when it is rather dark and a singing voice when it is relatively bright. With respect to the hue variable, our empirical choice of musical instruments is:

Figure 3-5. Association of instruments sounds with hue from red to purple in 360°.

- Oboe for red ($0 \leq H < 30$)
- Viola for orange ($30 \leq H < 60$)
- Pizzicato violin for yellow ($60 \leq H < 120$)
- Flute for green ($120 \leq H < 180$)
- Trumpet for cyan ($180 \leq H < 240$)
- Piano for blue ($240 \leq H < 300$)
- Saxophone for purple ($300 \leq H \leq 360$)

Note that for a given pixel of the sonified row, when the hue variable is exactly between two predefined hues, such as for instance between yellow and green, the resulting sound instrument mix is an equal proportion of the two corresponding instruments. More generally, hue values are rendered by two sound timbres whose gain depends on the proximity of the two closest hues. The audio representation h_h of a hue pixel value h is

$$h_h = g h_a + (1 - g) h_b$$

Equation 3-1. Proportional mixture of tow instruments in See ColOr.

with g representing the gain defined by

$$g = \frac{h_b - H}{h_b - h_a}$$

Equation 3-2. Gain of two mixed instruments in See ColOr.

with $h_a \leq H \leq h_b$, and h_a, h_b representing two successive hue values among red, orange, yellow, green, cyan, blue, and purple (the successor of purple is red). In this way, the transition between two successive hues is smooth.

The pitch of a selected instrument depends on the saturation value. We use four different saturation values by means of four different notes:

- Do (262 Hz) for ($0 \leq S < 0.25$);
- Sol (392 Hz) for ($0.25 \leq S < 0.5$);
- Sib (466 Hz) for ($0.5 \leq S < 0.75$);
- Mi (660 Hz) for ($0.75 \leq S \leq 1$).

When the luminance L is rather dark (i.e. less than 0.5) we mix the sound resulting from the H and S variables with a double bass using four possible notes depending on luminance level:

- Do (131 Hz) for ($0 \leq L < 0.125$);
- Sol (196 Hz) for ($0.125 \leq L < 0.25$);
- Sib (233 Hz) for ($0.25 \leq L < 0.375$);
- Mi (330 Hz) for ($0.375 \leq L < 0.5$).

A singing voice with also four different pitches (the same used for the double bass) is used with bright luminance (i.e. luminance above 0.5):

- Do (262 Hz) for ($0.5 \leq L < 0.625$);
- Sol (392 Hz) for ($0.625 \leq L < 0.75$);
- Sib (466 Hz) for ($0.75 \leq L < 0.875$);
- Mi (660 Hz) for ($0.875 \leq L \leq 1$).

Moreover, if luminance is close to zero, the perceived color is black and we discard in the final audio mix the musical instruments corresponding to the H and S variables. Similarly, if luminance is close to one, thus the perceived color is white we only retain in the final mix a singing voice. Note that with luminance close to 0.5 the final mix has just the hue and saturation components.

3.3.3 A relation between colors and sounds based on brain activity

The relation between colors and musical instruments has been typically studied by methods merely subjective, such as surveys and empirical matching conditions (see Sensorial color coding). Yet, an objective approach that permits a quantization of the existing relation (if any)

has largely been missing. In this thesis we posit a framework aimed at reaching this goal by co-relating EEG signals of a person, as follow: EEGs corresponding to the listening of sounds and the visualization of colors are collected. Then, a classifier to recognize the EEG pattern of one particular color is trained. Once trained, this classifier is tested using the EEGs related to sounds, in order to find an instrument eliciting similar EEG pattern. This idea rests on the assumption that we could identify functional fields in the primary auditory cortex involved in cross-modal transfer of the visual representation. The long-term objective here is twofold: Find out objective preferences for a particular individual, as well as general tendencies for color-instrument association into a population, if any.

The sounds of colors

With special focus on visually impaired assistance, quite a number of interfaces are these days attempting to transform visual cues into sound-based codes. Particularly, the association of color with music has drawn the attention of researchers in multiple areas. Yet, there are still many open questions about the optimal equivalence of colors when mapped to sounds. While this idea is broadly accepted, there are uncertainties on determining if a particular mapping turns out better than others when it comes to learning. Intuitively, one tends to link, for instance, bass sounds with dark colors or high pitched with visual brightness. Still, the relation of intermediate colors and sounds remains uncertain: With the eyes wide closed, does an electric guitar sound like red rather than blue?

Roughly speaking, two sorts of association methods can be found in the literature. Those relying upon subjective perception and others using empirical assumptions. On the one hand, the work presented by Rossi *et al.* [**138**] suggests a paradigmatic approach to the color-sound mapping that feeds on a broad survey. Thus, quite a number of perceptions are averaged forcing the subjectivity to collapse into a pattern. On the other hand, See ColOr [**29**] puts forward an empirical mapping of musical instruments onto the HSL color space (as seen in From colors to instruments sounds). The saturation variable (S) representing the degree of purity of hue (H) is rendered by sound pitch, while luminosity is represented by double bass when it is rather dark and a singing voice when it is relatively bright. Indeed, these assumptions are based on well accepted ideas in the musical community (at least in the western). Yet, instruments are mapped onto a 360° subdivision of H variable following no objective rules.

Beyond subjectivity

The idea of audio–visual cross-modal perception rests on the assumption that there must exist so-called polysensory areas [**144**], understood as brain areas activated by stimuli from more than one sensory pathway. A practical example to this is the well-known McGurk effect [**43**] that demonstrates the interaction between vision and hearing in perception of speech. This occurs when our brain hears a wrong audio cue for it is presented with visual evidence that something different is being said.

More generally, synesthesia [**145**] is regarded as a broader paradigm of cross-modal perception. Here, stimuli reaching one sensory/cognitive pathway automatically trigger involuntary reactions in a different sensitive pathway. This neurological condition can be manifested in various forms comprising spatial awareness, color and sound perception among others. Moreover, music is known to be a powerful elicitor of emotions, yet it is uncertain whether musically induced emotions are similar to those elicited by visual stimuli. This leads to an open question: "Can music-elicited emotion be transferred to and/or influence subsequent vision-elicited emotional processing?" [**146**]. In this spirit, Logeswaran *et al.* [**146**] showed preliminary evidence that such a cross-modal effect is indeed revealed (by event related brain potential components) at the very beginning of neuronal information processing.

Further, previous studies on EEG [**147**] have proved that on the one hand, positive musical stimulation results in a more pronounced lateralization on the left fronto-temporal cortices. And on the other hand, negative musical stimulation produces quite the same effect towards the right frontal areas. Likewise, such a frontal asymmetry is known to happen when affective visual stimuli are processed [**148**]. Consequently, overlaps between visual and musical perceptions are at least likely.

The proposed method

In a primary experiment, we aimed at mapping 10 colors onto 10 musical instruments sounds. A participant was tasked to carefully hear each sound during 40 seconds (see **Figure 3-6**). Subsequently, (s)he watched 10 colors sequentially. A 32-electrodes Biosemi [**149**] served to record the participant's EEG signals, for each sound played and each color watched. Thus, two data sets were collected $C_d=\{c_1,\ldots, c_{10}\}$ and $S_d=\{s_1,\ldots,s_{10}\}$ corresponding to colors and sounds respectively. Both C_d and S_d, can be regarded as a collection of 32-component patterns, whenever c_i and s_i are discrete signals of 32 channels each, $\forall i \in \{1,\ldots,10\}$.

Figure 3-6. A participant hearing sounds while his brain activity response to each sound (EEG signals) is recorded for further analysis.

Following, 10 Artificial Neural Networks $K=\{k_1,\ldots,k_{10}\}$ were used to classify the data belonging to C_d only. Therefore, k_i represents a binary classifier trained to assess whether or not a pattern belongs to c_i, $\forall i \in \{1,\ldots,10\}$ (i.e. this patter was caused by the i-*th* color). The use of neural networks (NN) allows us to determine a possible non-linear relationship between colors and sounds. A global sketch of this idea is shown in **Figure 3-6** as a framework to relate brain activity elicited by hearing sounds and watching colors [**150**].

Once the training is completed, we evaluated each classifier k_i using the data S_d. Our assumption here is that any pattern within S_d being classified positively by k_i, holds a close relation with those of c_i. Therefore, our aim is to identify the s_j with the highest number of patterns positively classified by k_i, $\forall i,j \in \{1,\ldots,10\}$. This finally leads us to establish a relation between the EEG elicited by the i-*th* color and that of the j-*th* sound. Further, we could assess the strength of this relation just by counting the number of patterns within s_j classified as positive.

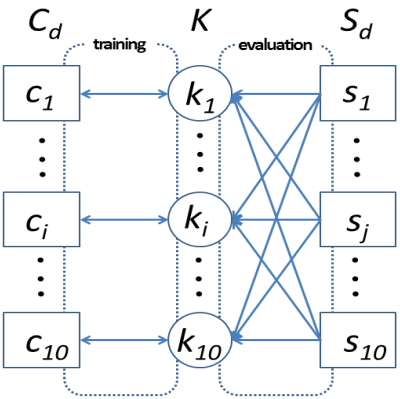


Figure 3-7. Proposed framework to relate brain activity elicited by colors with that of sounds. Though for the training of k_i all C_d data is actually used (c_i as positive examples and negative all c_j <$j\neq i$>). In this figure, we link k_i (in the training phase) only with the patterns it was learnt to recognize as positive (*i.e.* c_i). Oppositely, in the evaluation phase K is related to all S_d because evaluation entails all the sounds. After calculation of all these relations (*i.e.* the number of patterns within every s_i classified positively by k_i), we aim at finding the strongest relation in the evaluation phase.

Discussion

Our study on color and sound matching engages both, the visually impaired (no born blind) as well as, sighted individuals. Therefore, the acquisition of EEG related to colors may be achieved under two conditions: By watching steadily a flat colored screen, or simply by a sharp mental image of the intended color. In the future, we would like to find discriminatory features between these two populations. It is well known that blind people may eventually achieve greater hearing acuity; therefore, the visual-audio cross-model performance is likely to be further developed.

Another side of this work-in-progress that is worthy of attention, refers to the processing of the EEG signals. So far our analysis relies simply on time; we use a sliding window to average de 32-channels signal, so as to extract samples as patterns. Yet, typical EEG analysis is known to be more reliably performed on frequency domain. Our choice is based on the fact that larger number of patterns can be obtained so. A 40 seconds recorded signal may give a large number of samples in time, whereas in frequency domain only the very first frequencies are of importance. Nevertheless, in the future our results must be compared with those obtained in frequency so as to be validated.

Finally, at the end of this thesis analysis and gathering of data was still ongoing. The framework had also been implemented and tested in Matlab 7.0 using the Neural Networks Toolbox. Although preliminary results were promising, the evidence collected precludes drawing any assertions or generalizations due to the premature state of the work. Therefore, the contributions of this thesis in this regard consisted in:

- Put forward a novel and seminal idea that will permit an objective quantization of the existing relations (if any) in perception of colors and sounds.
- Modeling and implementation of a framework based on Artificial Neural Networks that allows evaluating the mentioned idea.
- Establishment of a protocol and gathering of preliminary data to be further studied by new PhD students.

3.3.4 How does See ColOr sound like?

Lateralization of sound

It is possible to simulate lateralization, also denoted as two-dimensional auditory spatialization, with appropriate delays and difference of intensity between the two ears. Nevertheless, inter-aural time delay (ITD) and inter-aural intensity difference (IID) are inadequate for reproducing the perception of elevation, which represents a crucial auditory feature for 3D spatialization. In fact, the folds of the pinnae cause echoes with minute time delays within a range of 0-0.3 ms [**151**] that cause the spectral content of the eardrum to differ significantly from that of the sound source. Strong spatial location effects are produced by convolving an

instrument sound with head related impulse responses (HRIRs) [**151**], which not only varies in a complex way with azimuth, elevation, range, and frequency, but also varies significantly from person to person [**152**].

Generally, reproducing lateralization with uncustomized HRIRs is satisfactory, while the perception of elevation is poor. Since one of our long term goals is to produce a widely distributed prototype, thus involving standard HRIRs, we only reproduce spatial lateralization of 25 points through the azimuth plane with the use of the CIPIC database [**68**]. The effect caused to the user is the perception of sounds coming from 25 different fontal locations (virtual sources), from left to right (see **Figure 3-8**). In practice, each sonified point corresponds to the convolution of an instrument sound (300 ms) with the corresponding HRIR filter related to a particular azimuth position. We only reproduce spatial lateralization with the use of the HRIR measurements belonging to the CIPIC database [**67**]. This database is one of the most used for high-spatial-resolution measurements. Release 1.0 was produced for 45 different persons; specifically, impulse responses were produced in quite a number of distinct directions for each ear and for each subject.

A particular HRIR can be evaluated after calculating a Head Related Transfer function (HRTF) [**67**]. The HRTF is a response that characterizes how an ear receives a sound from a point in space. In other words, describing how a sound from a specific point will arrive at the ear can be modeled using the HRTF. Usually, practical simulation of HRTF is achieved with the use of a dummy head (**Figure 3-8**) equipped with small microphones at either ears. Sounds emitted around 369° degrees are recorded by the microphones. These records serve to support a mathematical modeling of the HRTF.

Figure 3-8. Spatialization of sound in See ColOr. Sounds can be perceived as approaching from 25 different positions (from right to left) across the azimuth-frontal auditory field.

The sound of local and global modules

In the local module of See ColOr, the sonified part of a captured image is a row of 25 pixels relative to the central part of the video image. As replicating a crude model of the human visual system, pixels near the center of the sonified row have high resolution, while pixels close to the left and right borders have low resolution (peripheral vision). This is achieved by considering a sonification mask indicating the number of pixel values to skip. As shown below, starting from the middle point (in bold), the following vector of 25 points represents the number of skipped pixels:

[15 12 9 7 5 3 3 2 2 1 1 1 **0** 1 1 1 2 2 3 3 5 7 9 12 15]

Equation 3-3. The mask of resolution for sonification of 25 pixels relative to the central part of a captured image in the local module of See ColOr. The number (0) in the middle of the mask (in bold) represents the central pixel of the image. The others numbers in this sonification mask must be understood as the number of skipped pixels. This masks mimics a crude model of the human visual system, pixels near the center of the sonified row have high resolution, while pixels close to the left and right borders have low resolution

These 25 pixels are sonified simultaneously using lateral spatialization from left to right (Lateralization of sound). Thus, the central part of the image is mapped (and augmented) into the entire azimuth-frontal auditory field (left side of **Figure 3-8**). Accordingly, this module was named local, since it focuses exclusively on a row of 25 pixels localized in the center of the image. And though we try to mimic peripheral vision using **Equation 3-3**, this only applies to that local portion of the image. This fact, unfortunately, gives rise to an important drawback in See ColOr that is often liken to the tunneling vision effect (**Figure 3-9**) in retinitis pigmentosa [**105**]. In principal, this problem dramatically affects the understanding of an image (as a whole) because peripheral vision or global perception is unattainable. To make it worse, the small perceptible portion of the image is hardly understandable as it is devoid of context. To cope with this downside the global module was created.

Figure 3-9. Tunneling vision phenomena. In the right image the field of view is so narrow that achieving the general aspects of the image (in the left) is unlikely. Further, the only

visible part is hardly understandable as it lacks for context. The tunneling vision is a progressive effect of retinitis pigmentosa, a degenerative eye disease responsible for great percentage of blindness.

The global module[10] sonifies the color of a pixel tapped with the fingertip on a tablet where the image is mapped (see **Figure 3-10** right side). Therefore, only one source is heard in this module (or as many as fingers tapping on the tablet). Here, the image (or field of view) is divided into 25 equal parts, so sources may be heard from 25 different positions in the image, according to the finger (see **Figure 3-11**). The fact that we do not sonify elevation, however, is reflected in **Figure 3-10** by displaying a source that matches the finger horizontally, though not in elevation. In other words, the virtual sources are always listened on the azimuth plane regardless the vertical coordinate of the selected pixel. 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. This is mostly the case because (s)he can access various points simultaneously (several fingers), slide a finger so as to scan areas and so forth.

Figure 3-10. (left) The sonification in the local module is illustrated. There are 25 points and 25 sources in this module. To effects of visualization however, only 3 points and 8 sources 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

[10] Importantly, the audio-output of this local module does not lie into the soundscape category, for two reasons: Firstly, its complexity is fairly low, as it sonifies only 25 out of 640x480 image pixels. Secondly, those 25 sonified pixels provide information only about the small central part of the image, opposite to soundcapes which are intended to provide a panoramic of the picture.

auditory field. (right)An illustration of the sonification in 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), which is why in this illustration the source matches the position of the point horizontally though not in elevation. In other words, the source is put down on the azimuth plane, preserving only the horizontal position of the finger on the image. It is well known that rendering elevation is much more complicated than lateralization.

Finally, while spatialization of sound may reflect the lateral location of a virtual source. The distance of this source with respect to the user is assumed as the depth of the selected pixel. As an example, if we locate a virtual source in **Figure 3-11** on top of the tree, the emitted sound will be perceived as coming from the third spot and as far as the depth of the pixel. In both modules of See ColOr global and local, depth is represented by sound duration and the mapping is given by:

- 90 ms for undetermined depth;
- 160 ms for ($0 \leq D < 1$);
- 207 ms for ($1 \leq D < 2$);
- 254 ms for ($2 \leq D < 3$);
- 300 ms for $D > 3$.

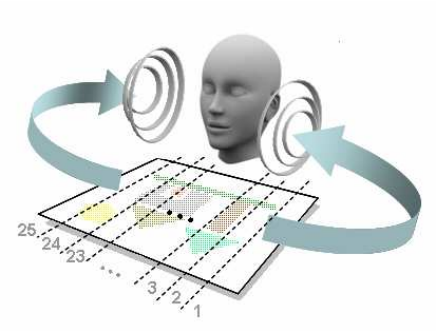


Figure 3-11. In the global module, the field of view is partitioned into 25 parts from within which virtual sources may be placed.

Consequently, the depth of a point will be perceived as the rhythm of the sound: The closer the point, the faster the rhythm. Note that it is possible to have points of undetermined depth, especially in areas occluded, for which the depth camera is unable to determine parameters related to the calculation of the disparity between the left and right images. To conclude, as sounds in See ColOr originally last 300 ms, they need to be cut off in accordance with depth. However, cutting off an audio signal by removing abruptly the undesirable part

produces a sudden cutting sound (i.e. a sharp click in the end). This turns out annoying as the user listens to the flow of sounds. In order to avoid the undesired effect, we use a smoothing filter of 22.5 ms to cut off the signal at needing. Thus, 22.5 ms before the signal stops, it begins to lose amplitude gradually (see **Figure 3-12**).

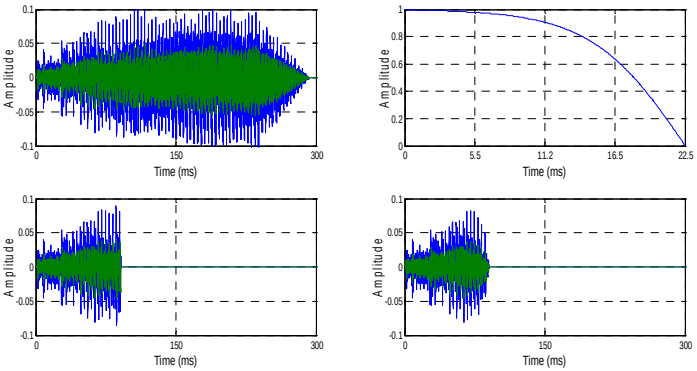


Figure 3-12. On top left we have a typical sound signal used in See ColOr. Top right is a smoothing filter that multiplies 20 ms of the signal at certain time determined by depth. In the bottom left figure, the signal has been cut off at 90 ms without a filter. Finally, bottom right figure shows the same signal as it is cut off smoothly. Notice that within its last 20 ms the final signal gradually loses amplitude in order to avoid a sudden cutting sound (i.e. a sharp click).

3.3.5 Acoustic virtual objects

Awareness of entities in the environment is essential for assisted navigation. The goal here is to sonify objects that do not intrinsically produce sound, with the purpose of revealing their nature and location to the user (i.e. acoustic virtual objects [**8**]). Although, auditory cues are regarded as an excellent approach to representing objects, it is important to identify methods or techniques that can guarantee learnability. For instance, here we face once again the paradigm of using natural language: spelling out the name of the objects seems a simple yet efficient solution. Accordingly, Dodson *et al.* [**153**] assume that 'since a blind human is the intended navigator a *speech* user-interface is used to implement this'. Nonetheless, many authors such as Franklin *et al.* [**154**] argue otherwise. At large, they claim that having a voice announcing the object leads to difficulties in interpreting spatial relations from common speech. Another reason is that processing speech requires a lot of mental resources and also, the lengthy spelling of names might reduce the information transfer rate. Last but not least, the use of natural speech introduces language dependency. Essentially in this stage of this

thesis we want to explore alternatives to the use of common speech and further, we wanted to verify objectively the advantages of this method in comparison with others.

Methods for object sonification others than speech

A number of alternatives to the use of natural speech have been proposed in the literature. The main focus of this research is the learnability of sound cues representing environmental features. Though, accessibility to main stream technologies is another field in which sonfication of entities is used to improve navigation in auditory menus. In principal, the tendency to use natural speech in this problem seems intuitive, however, it might be not necessary optimal. As an example think of text for describing an entity in visual contents. Though intuitive, this very often turns out to be less efficient than icons: Hemenway [**155**] said that icons are fairly more easily processed and located than words, because meaning is derived right from the object or action they represent. Further, Kolers *et al.* [**156**] show how to reach beyond linguistic and cultural boundaries with the use of icons.

In this view, one of the most popular approaches nowadays is the ***Auditory icon*** [**157**], a brief audio cue that represents the intrinsic sound of an object. Just as several dimensions such as shape, color, etc. can be encoded within a visual icon and then processed in parallel by sight, so does the auditory pathway (pitch, amplitude, timbre, etc.) with respect to audio icons [**157**]. The idea is simple, the sound of a car represents a car as well as a flushing toilet stands for restrooms. The idea loses its simplicity; however, if we think that while direct relations are inferred from the sound made by the target, indirect relations also can arise with surrogates for the target. For instance, the sound of an engine may represent a car though also the engine itself, or any other machine that needs an engine for power supply. To make it worse, what could be the sound to represent a table, a t-shirt or a hat? As long as a sound evokes unequivocally the associated object, auditory icons come in handy. Otherwise, their utility in representing entities is severely limited, especially for representing ambiguous concepts or objects producing no sound intrinsically.

In sight of this, the authors in [**158**] propose ***Earcons*** as an alternative to the use of natural speech and auditory icons. Earcons are abstract, synthetic and mostly musical tones or sound patterns that can be associated indistinctively to objects. Since there is no need of an intrinsic relation between the audio cue and its associated object, earcons do not suffer from any of the problems affecting auditory icons. In fact, the use of earcons is limitless and extendible to objects, actions, concepts and so forth. A typical example of an earcon is the sound of Windows being started in a computer. An earcon is very often regarded as the auditory counterpart of a company's icon; this is when a company is recognized by a synthetic sound. When one sees a company's icon you will instantly know who they are, the 'earcon' works in the same way but when one hears the audio cue one immediately thinks of the company it is connected with. A key aspect of earcons the hierarchical property they have to form families of sounds. Once, you have an earcon for a chair, for instance a red chair, a plastic chair, an office chair etc. can be associated to the same earcon with modifications of pitch, rhythm,

timber and so forth. An important drawback of earcons refers to learnability since these sounds present no natural relation with their associated items. Back to the Windows example, it has taken years for users to be familiar with it. In other words, training is an issue when it comes to earcons.

Lately, ***Spearcons*** [**159**] appeared as a promising alternative to the use of classical methods for audio representation of objects. They may be regarded as a further speech-based type of acoustic representation [**159**]: Spearcons use spoken phrases that are speeded up to the point they may no longer be recognized as speech (i.e. they are not fast speech). Here the accelerated version of the spoken object's names produces sounds alike earcons. Nevertheless, because the mapping between objects and spearcons is non-arbitrary, less learning is expected. The spearcon associated to an entity is unique because it is acoustically related to the underlying speech phrase. However, objects of the same type are expected to produce similar spearcons as they are rooted in a common word (e.g. desktop red, desktop green, desktop black; the sound of the word desktop is the root). Thus, one can say that spearcons are distinct but in the same time, present hierarchical associativity. Therefore, much like earcons, spearcons have the same capacity to form families of sounds. Broadly, spearcons are analogous to a fingerprint, a unique identifier that is only part of the information contained in the original. In **Figure 3-13** the problem of audio representation of objects using the here studied methods is posited.

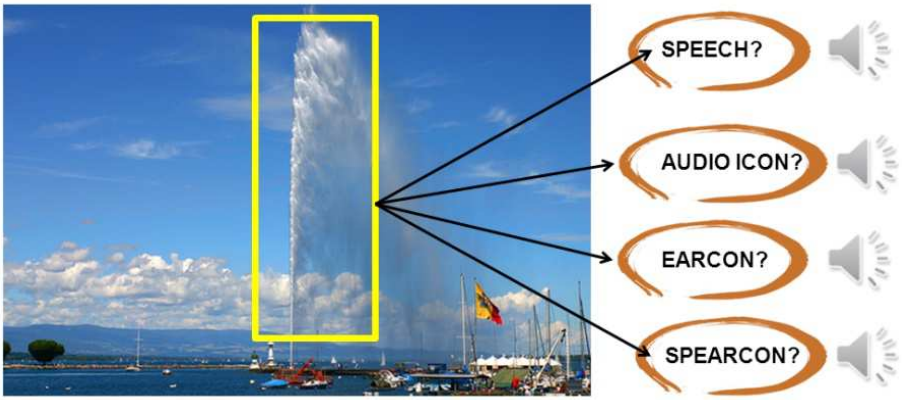


Figure 3-13. In this figure the dilemma of audio representation for water is illustrated: A running water sound, a voice saying the word 'water', an earcon, a spearcon. What's better?

Experiments with acoustic objects

As discussed, the sonification methods mentioned earlier have all advantages and disadvantages and are relatively common in auditory displays. When it comes to visually impaired assistance, however, the main concern is how learnable and how accurate the method is, as this will affect the overall usability in real scenarios. We conducted a comprehensive study with visually impaired individuals to investigate the usability of these methods (see **Figure 3-14**). Our findings (see **Figure 3-16** and **Figure 3-17**) suggest that while natural speech is apparently the most promising approach; spearcons are undeniably advisable if we consider language-dependency and all concerns associated with speech in earlier sections.

Figure 3-14. Some participants taking this experiment on acoustic virtual objects.

Fifteen graduate students who reported visual acuity of less than 20/400 participated in this study. There were 9 male and 6 female students, who ranged in age from 18 to 40 (see **Figure 3-14**). Seven objects were represented using the three categories of sonification described in the previous section as well as natural speech (elements and categories are presented in **Table 3-1**). For each category, participants were taught the meaning of each audio cue. In other words, the training comprised the presentation of the object or entity, followed by its associated audio cue. In each modality, however, the seven objects (and their auditory representations) were presented as many times as required by the participant (amount of training). A series of audio lists was created then, seven lists per category (7x4=28 lists). Firstly, we formed four lists (category 1 to 3 and speech) in which the position of object 1 was randomly assigned though preserved across the four lists. We repeated the same process for the 6 remaining objects, so that we obtained seven sets of four lists (one set of four per object) with elements sorted equally between categories but not between sets. All these lists were then put into a repository with no order criteria. This process is illustrated in **Figure 3-15**:

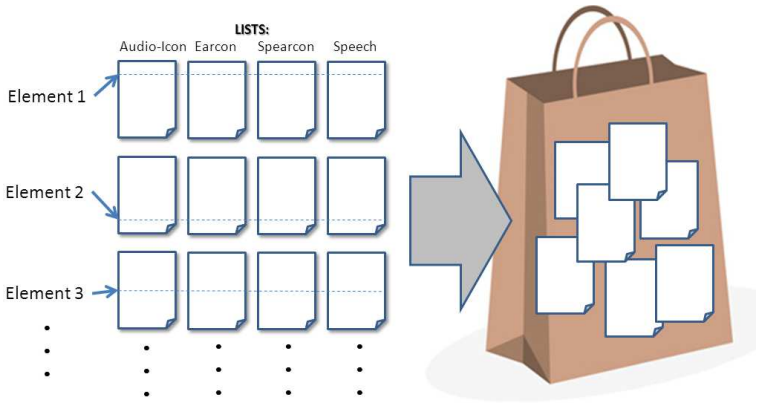


Figure 3-15. Preparation of audio lists for the test on object sonification methods.

Participants were tasked to hear all the lists within the repository. They could go through the audio elements of a list using the right click of the mouse (forward) or the left one (backward). In this way participants could hear a cue as many times as needed. Also, there was no need to hear the entire audio element or cue, so participants could advance faster or slower in the search of a specific element just by clicking. The lists were circular, so that from the bottom, a click took the user to the top. If the target was found, the search terminates by a middle-button click. Ultimately, we wanted the participants to find each object as presented in four different modalities, though at the same position. Here, the position is relevant as it guarantees the same conditions for a sought object: we cannot compare times in finding an object within the list of spearcons and speechs, if such an object is on top of the former and in the bottom of the latter.

	Auditory Icon	Earcon	Spearcon	Speech
Bell.wav	4 seconds	0.3 seconds	0.6 seconds	1 seconds
Bird.wav	4 seconds	0.3 seconds	0.6 seconds	1 seconds
Car.wav	5 seconds	0.3 seconds	0.6 seconds	1 seconds
Snoring.wav	6 seconds	0.3 seconds	1.2 seconds	1.6 seconds
Train.wav	4 seconds	0.3 seconds	0.8 seconds	1.2 seconds
Water.wav	5 seconds	0.3 seconds	0.9 seconds	1.3 seconds
Woman.wav	4 seconds	0.3 seconds	1.2 seconds	1.6 seconds

Table 3-1. Elements (objects or entities) and methos of sonifcation. This table also registers the duration of each audio cue (acoustic object) in reliance with the method. To create earcons we used the some of the instruments sounds from See ColOr.

Since the lists were labeled as they were created, we knew what object (out of seven) to ask the user find as he picked a list from the repository (**Figure 3-15**). The lists were picked from a randomly-formed repository, instead of presented in order just to preclude transfer learning. For instance, after being asked to find the first element (object) within the audio-icon and spearcon lists, the participant might learn the position of that element since it is preserved between categories. Finally when the participants processed the 28 lists, we repeated the experiment 10 times. Therefore, percentage of accuracy and times are averaged in the results reflected below (**Figure 3-16** and **Figure 3-17**).

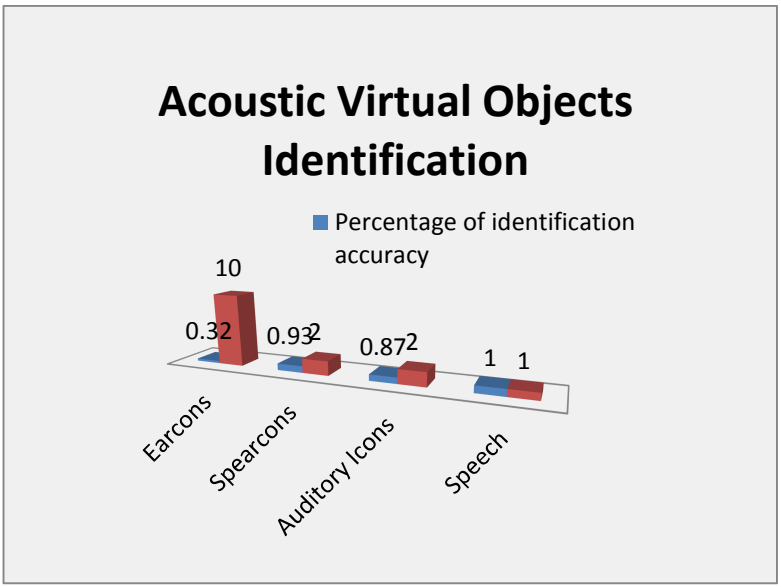
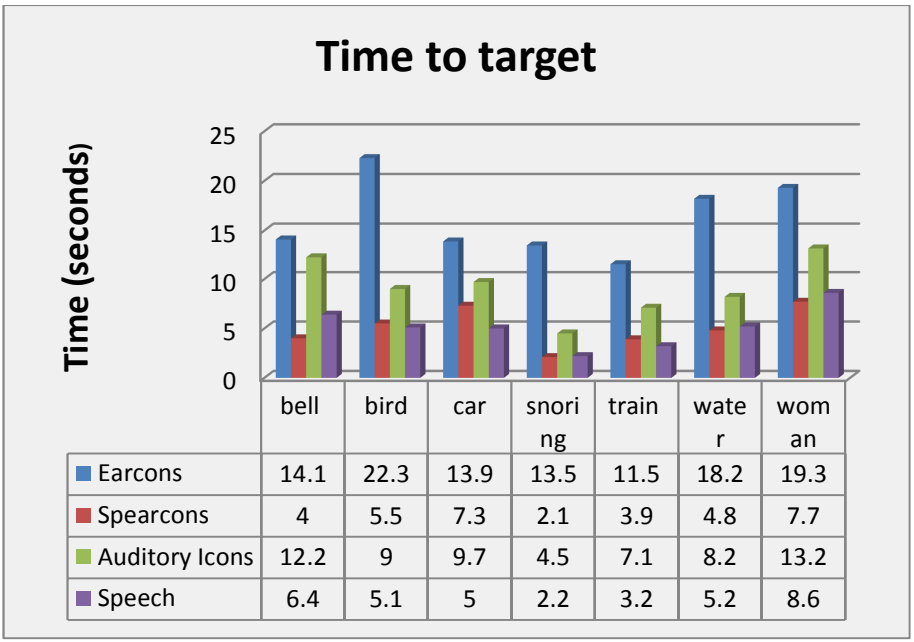


Figure 3-16. Training time and testing accuracy in this experiment. Training is given once at the beginning of the experiment, and accuracy is averaged from the 10 times that the participants went through the 28 lists.



	bell	bird	car	snoring	train	water	woman
Earcons	14.1	22.3	13.9	13.5	11.5	18.2	19.3
Spearcons	4	5.5	7.3	2.1	3.9	4.8	7.7
Auditory Icons	12.2	9	9.7	4.5	7.1	8.2	13.2
Speech	6.4	5.1	5	2.2	3.2	5.2	8.6

Figure 3-17. Time needed to identify an acoustic virtual object using four methods of sonification. Note that these numbers are averaged from the 10 times that the participants went through the 28 labeled lists. It is also worth noticing that the time to find a particular object, in principle, is based on its position within the lists. Therefore, in this figure only comparisons of times between methods are valid and not between objects.

Earcons are unadvisable at all: they need more training, present the lowest accuracy and they are the most delayed to find. Notice that while generally showing good accuracy, auditory icons take more time to be found than spearcons and speech. This is because an auditory icon in average lasts 4.5 seconds, whereas speech lasts 1.1 seconds, and some less a spearcon (300 ms, see **Table 3-1**). Moreover, in terms of training/accuracy no noticeable difference was found between auditory icons and spearcons, while one lacks for accuracy, the other requires more training. Nevertheless, in regard to the time for reaching a target the spearcon is a much faster technique. Compared to spearcons, speech needs moderatly less training, and its accuracy is slightly better. However, not necessarily the time needed to reach a target is shorter with the speech, occasionally the spearcon does better. By and large, our findings suggest that while natural speech is apparently the most promising approach; spearcons are quite a good strategy if we consider language-dependency and all concerns associated with speech in earlier sections. The use of spearcons could be moderately advisable. Specially, language dependency will reduce the potential population that See ColOr aims at reaching, or

will introduce the need of re-adaptation and re-implementation of the system for each country or community.

3.4 Efficient registration of range and color images

3.4.1 Description

With the increasing use of 3D entertainment and multipurpose representation of virtual environments, range cameras continue to gain in popularity as prices are getting lower. While generally promising, there are shortcomings to the use of these sensors, which need to be resolved. Particularly, these cameras lack for color and some do not even provide a grey-level or intensity image. This fact dramatically diminishes their usability in See ColOr due to fundamental reasons: Color is essential for the global and local modules, as well as intensity images are needed for expansion of See ColOr into computer vision (i.e. perception module). As noted previously in this thesis, earlier prototypes of See ColOr used cameras such as, PMD [vision] ® CamCube 3.0, or Bumblebee™ - Point Grey Research. Therefore, without efficient coupling of color and range images the evolution of See ColOr would have been impossible.

The advent of Microsoft Kinect (a cost-efficient solution) partly alleviated this shortcoming by embedding a depth-color camera pair in one sensor. Unfortunately, Kinect's internal color camera often lags behind the needs for quality in mainstream applications. Moreover, even if there is no external camera being added, yet the internal sensors (depth and color) of Kinect do need to be registered. Remember that first Kinect cameras in the market, before SDK were launched, did not provide calibration algorithm whatsoever. In such a context, the use of an external HD color camera (and in general color-depth registration) began to draw our attention in regard to See ColOr. It is worth noticing that beyond See ColOr, coupling range and HD-color cameras benefits a broad range of applications in which neither alone would suffice.

Although a number of interesting ideas emerged from this problem, when it comes to couple two camera systems, image registration is perhaps the most affordable approach. Yet, classic registration methods yield no suitable results in this particular case. Much is known about intensity images registration; however there are still many open questions about registering an intensity image and a surface that lacks color and geometric features. In this spirit, the aim in this stage was to produce a general method to register range and RGB-digital images (see **Figure 3-18**):

We present here a simple yet highly efficient method to register range and color images. This method does not rely upon calibration parameters nor does it use visual features analysis. Our initial assumption is that the transformation that registers the images is a linear function of the depth. Drastically enhanced performances in the computational processing are attained under this condition. Nonetheless, the linearity assumption is not met for cameras others than Kinect. We show, however, that ultimately this method is independent of the mathematical model underlying it (be it linear or not). Therefore, the efficiency of this ap-

proach is preserved entirely in both cases. Further, this section reports on the results of experiments conducted with various range camera models that endorse the proposed method. Eventually, three key features can be derived from this technique: practicality, accuracy and wide applicability. Although, in principle this method presents limitations relating mainly to image distortion in some other sensors tested, we also study how to cope with this drawback with no loss of efficiency. In fact, this method might be regarded as an approach to correct distortion in range images, an issue that remains challenging.

Figure 3-18. Coupling of color and range cameras: Leftmost column shows two mounted systems made up by depth sensors (Kinect on top, SwissRanger SR-400 at the bottom) and a HD web camera. In the middle column the re-sized depth maps and the color images (webcam) have been merged. In the rightmost column finally, we repeat the merging right after the depth maps have been registered into the color images using our algorithm. A twofold aim may be targeted: the addition of depth in a HD cam or the improvement in color resolution of kinect. Moreover, images at the bottom of this figure may attest that our algorithm is valid for complex scenes which exhibit flexural[11] geometry.

[11] the state of being flexed

3.4.2 Previous approaches

The registration of RGB and range images of a same scene aims at matching color images and surfaces which lack color [**160**]. This problem remains largely unexplored in computer vision. Nonetheless its applicability is fairly well defined. As examples it is worth mentioning: 3D-laser extrinsic parameters estimation [**161**], color improvement in depth-color camera pairs [**162**], and, joint-depth and -color calibration [**163**]. Particularly, extrinsic calibration of colorless ToF (Time-of-Flight) cameras or 3D-lasers is a relentless challenge that is usually approached in more refined ways: Hirzinger *et al.* [**164**] describe a multi-spline model that requires a robotic arm to know the exact pose of intended sensor. Zhu *et al.* [**165**] describe a high-cost algorithm for fusing stereo-based depth and ToF cameras via triangulation. Unfortunately, a method easy-to-use, accurate, and applicable to a wide range of sensor has largely been missing.

An assessment of the general problem of image registration might be useful. In general, a vast range of techniques exist in the literature. Yet, more needs to be done to progress toward general solutions, if any. In 2003 Zitova and Flusser [**166**] published a complete review of classic and recent image registration methods. Following, Deshmukh *et al.* widened the spectrum of solutions by including updated advances in a more recent review in 2011 [**167**]. In all these works the image registration problem is defined as the matching of images of a scene taken from different sources, viewpoints and/or times. Yet, the former condition (inter-source) is limited to the variability of RGB sources only. Therefore, registration methods such as the one proposed by Hsinchu *et al.* [**168**] using artificial neural networks, or others that use belief propagation strategies as is the case of Sun *et al.* in [**169**], are likely to fail. This is mostly the case because they rely on the matching of color-based visual features common (or mappable) in both images.

In mainstream applications of computer vision, depth and color together as complementary cues about the scene are highly desirable [**163**]. Yet, while low resolution of the ToF camera is enough to segment depth-based areas, higher resolution RGB camera allows for accurate image processing. In this spirit, Huhle *et al.* [**160**] present a novel registration method that combines geometry and color information in order to couple a PMD (photonic mixer device) camera with an external color camera. The alignment carried out in this work is based on the Normal Distributions Transform (NDT, [**170**]) of the range images and a Scale Invariant Feature Transform (SIFT, [**100**]) feature detector applied to the high-resolution color images. Thus, the authors claim to combine the robustness of the globally applicable feature-based approach and the precise local fitting via NDT. More recently, in 2011, Van Gool *et al.* [**162**] combined a digital camera and SwissRange ToF sensor using a regular calibration method for stereo systems [**171**]. The key idea of this approach was treating the output of the range sensor as though it was a RGB image.

In [**172**] the authors conduct a comparative study of some of the most important Depth-and-color calibration algorithms. This work includes implementations as well as performance

comparisons in both, real-world experiments and simulation. Two algorithms stand out in this study: first, Zhang's method [**173**] that presents a maximum likelihood solution. This method uses checkboards and relies on co-planar assumptions. Also, manual correspondences need to be specified to improve calibration accuracy. Second is Herrera's method [**174**], whose authors claim to achieve features such as accuracy, practicality, and applicability. The method requires planar surface to be imaged from various poses and presents a new depth distortion model for the depth sensor. Although, one more method called DCCT (Depth-camera calibration toolbox) is studied in [**172**] [**175**], the article about this method is to date unpublished. Also, the authors in [**173**] have not shared their code, so that we only use [**174**] for comparisons later in this work.

Finally, as discussed by Han *et al.* in [**176**], the "parallax" is perhaps the most challenging problem when it comes to image registration. Algorithms suffer from this problem by virtue of the assumption that the scene can be regarded approximately planar. This is of course not satisfied by large depth variation in the images with raised objects [**177**]. Paulson *et al.* [**176**] presented in 2011 an outstanding idea to cope with the parallax problem by leveraging approximated depth information. Basically, their idea was to recover the depth in the image region with high-rise objects to build accurate transform function for image registration. The drawbacks from which this method suffers are fourfold: motion camera parameters are vital, significant manual work is needed, inaccurate approximations based on heuristics are very likely, no real time. It is worth noticing that by feeding from a depth source (ToF sensor) the parallax phenomenon is no longer an issue in our method.

3.4.3 A new strategy

Background

An image is in theory an infinite assemblage of successive planes that eventually makes up depth effect. Thus, in stereo-vision systems (stereo images captured by a camera rig) depth is discretized into many parallel planes (see Figure **3-19**). The shift required to attain an exact overlap of two parallel planes is well known as the disparity [**178**]. Disparity is usually computed as a shift to the right of a point when viewed in the left plane (distance between blue 'left' and red 'right' points in Figure **3-19**). Also in Figure **3-19**, we can see that each pair of parallel planes presents a different amount of disparity (e.g. parallel planes captured at d_1 and d_2). Furthermore, the following observations may be made on the same figure:

a. Only the *x* axis is prone to have disparity.
b. The disparity decreases as the distance of the planes (d_i) augments.
c. Disparity is constant for all the points into parallel planes.

It is worth noticing that (a) shall be met provided that the stereo images are first rectified. Thus, both images are rotated (until their epipolar lines get aligned [**178**]) to allow for

disparities in only the horizontal direction (i.e. there is no disparity in the *y* coordinates) [**65**]. Regarding (b), while the relation there pointed might be a common place observation; the rate at which it is given will be further studied in this section (i.e. disparity vs. depth). Within the next sub-sections it will be shown that though this relation is non-linear, exceptions can be made in this regard (Kinect's case). Finally for (c), we want to stress that this is very much expected in an ideal system, reason why we hypothesize about it. Nonetheless, practical experiments conducted later in this section will contribute to make this assumption clearer.

Ultimately, actual registration of two images demands a functional description of the displacements (disparities) between parallel planes across the depth. Thus, objects lying on an image plane (left) shall be accurately shifted to their counterparts on the parallel image plane (right). Our idea is to sample as many pairs of points as possible (blue-red pairs in Figure **3-19**) in as many parallel planes as possible too. Thus, we can interpolate the function that describes twofold information: firstly, the variation of the disparity between parallel planes, if any (it is thought to be constant by far). Secondly, the variation of the disparities with depth (it is expected to be linear for Kinect). Before we go any further with this idea, however, some important aspects need to be studied in order to endorse the assumptions made so far. This will be of help later in formulating our algorithm.

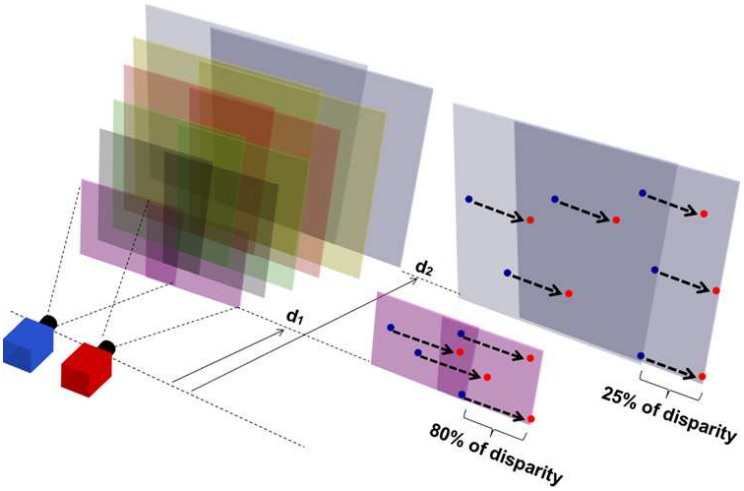


Figure 3-19. Disparity and depth planes in a camera rig. Parallel planes of the images increasingly overlap each other with distance. Two parallel planes need a constant shift (disparity) to fully overlap (matched). Under ideal conditions, this disparity must be constant and decreases with depth

Depth vs. Disparity

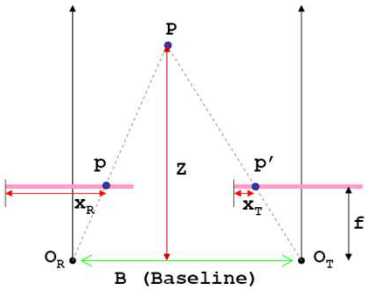


Figure 3-20. Stereo system set up. Orthogonal projection of Figure 3-19, also known as standard stereo configuration: P is a point in the 3D space whose depth (Z) may be recovered using p and p' (its projections into the focal planes placed at f). B is the distance between O_R and O_T (the cameras). As long as the system has been rectified, the disparity may be assessed by subtracting x_R and x_T (the x coordinate values for p and p').

In **Figure 3-20**, an upper-view of the standard stereo configuration with rectified images is presented. When aiming at recovering the position of *P* (a point in the space) from its projections *p* and *p'*, we need to consider similar triangles (ΔPO_RO_T and $\Delta Ppp'$):

$$\frac{B}{Z} = \frac{(B + x_T) - x_R}{Z - f} \Rightarrow Z = \frac{Bf}{x_R - x_T} = \frac{Bf}{d} \Rightarrow Z(d) = \frac{Bf}{d}$$

Equation 3-4. Depth (Z) in function of disparity (d)

where x_R-x_T is the disparity (*d*), *Z* is the depth of *P*, and *B* represents the distance between the two cameras. The fixate location (*f*) is known as the distance in which the planes of projections are fixed.

Notice that in order to substitute one of the cameras (either O_R or O_T) by a depth sensor in **Figure 3-20**, few considerations are only needed: The range map is to be regarded as a regular image within which disparities with its colored peer may be encountered. Also, *Z* turns into a known variable accessible from the range map itself. With this in mind, **Equation 3-4** still holds when a color camera is replaced. It needs to be said, therefore, that the function describing the relation 'disparity (*d*) vs. depth (*Z*)' is non-linear, though our hypothesis argues otherwise.

In this work, however, it will be shown that the Microsoft Kinect is a sensor for which disparity can be modeled as a linear function of depth within its depth range. M.R. Andersen

et al. [**179**] have showed the linearity of Kinect through experimental results from which they have concluded: "*The raw measurements provided by the sensor are non-linear, but the data are linearized in the OpenNI software*". Besides, we argue that **Equation 3-4** might very well be approximated by a linear function for values of Z lying within a reduced domain. It is well known that the effective field of view of Kinect is rather small (see **Figure 3-21**.) compared to other range cameras. Tilak Dutta summarizes the operational volume of this camera in [**180**] as follows: "*the effective field of view finally decreased to 54.0° horizontally and 39.1°*". This effective field of view corresponds to the 3D measurement volume shown in **Figure 3-21**. Finally and anyhow, we will present next experimental results that support this linearity for the Kinect sensor. Cases of range cameras for which this linearity does not hold, will be addressed later in this paper.

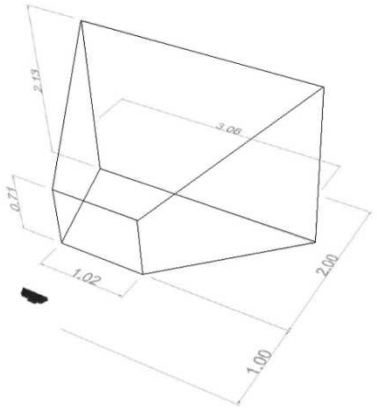


Figure 3-21. Kinect effective field of view: *3D measurement volume of the Kinect sensor.*

Kinect

The schema depicted in Figure **3-19** has been applied on a Kinect-provided image pair (color image and depth map) as follows: The disparity was randomly sampled within parallel planes, for a number of depth levels (a pair of parallel planes per level). **Figure 3-22** renders such a procedure for a particular point (P) in a pair of images. Further, **Figure 3-23** plots all sampled disparities as a function of both, the image coordinate (x, y separately) and the depth. This figure reveals twofold information: On the one hand, while the disparity is not constant into parallel planes (which argues against our expectations), it does vary linearly. On the other hand, identical behavior can be observed through depth (i.e. the disparity decreases also linearly as depth augments). Following these observations, planes have been used to fit the data shown in **Figure 3-23**. Nonetheless, experimental results shown in next

sections will be of help to prove that in fact, when it comes to Kinect, it is a plane the model that interpolates the best.

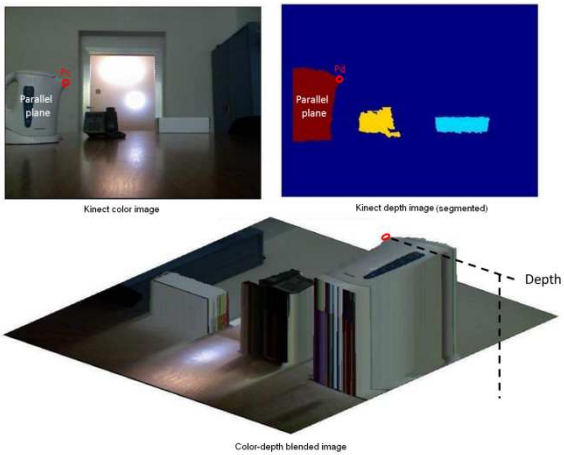


Figure 3-22. Depth color blending: Pc represents a point given in a color image, as well as Pd represents its counterpart in the range image. By measuring the distance (in pixels) between Pc and Pd, a sample of the disparity between the parallel planes (given at "Depth"), can be assessed. Notice that for this figure color and depth images were already calibrated (depth map was modified) and merged on the bottom, therefore, the disparity was corrected to zero.

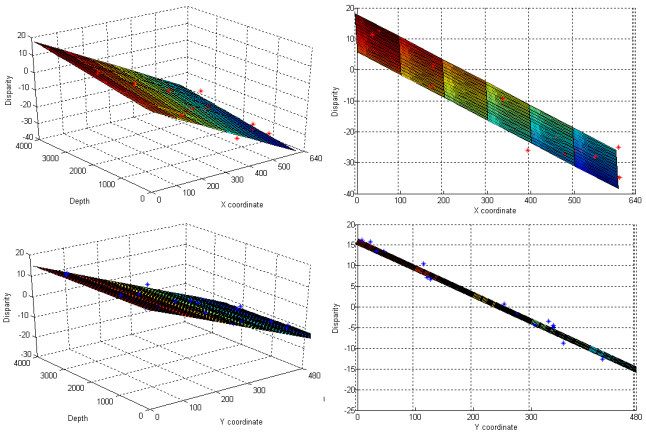


Figure 3-23. First column of these four plots represents the disparity of a pair of kinect-provided images (depth-color) as a function of its coordinates x (top) and y (bottom). The

second column shows a disparity-coordinate view of the same functions where the linearity of the data becomes evident.

It is worth noticing the fact that having a disparity distinct to zero across the *y* coordinate, indicates that the stereo images lack for rectification. Moreover, sampled disparities do not rest exactly over the planes due to the manual markup, which introduces an error (± 3 pixels). When depth maps are intended as regular images, these errors are very likely (e.g. boundaries are not reliable due to physical issues of range measurement hardware). Therefore, measuring the disparities between a range and a color image is a task that demands human intervention. See **Figure 3-24** where it is shown that sometimes, along with the zoom-in, eye-based extrapolation is a need for matching two points and evaluate their disparity.

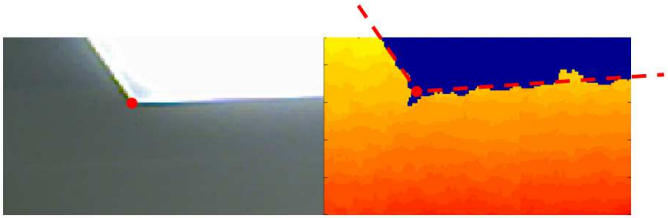


Figure 3-24. A point (red) manually marked within a zoomed area in both: a color image (right) and a depth map (left). While manual markup to the right is regarded as an easy task, manual markup to the left is not. Automatic algorithms fail to detect this point due to boundaries' discontinuities caused, in turn, by physical issues of range measurement hardware. The dashed lines represent the human-eye intervention needed to approximate the boundaries and calculate the point precisely.

Finally, in order to verify that the 20 points plotted in **Figure 3-23** obey a uniform distribution across a plane surface, the following test was conducted: 100 triplets of points were randomly selected and for each triplet the orthogonal vector was calculated using vector products. This test revealed that orthogonal vectors arising from the data diverge by negligible extent. Furthermore, the mean orthogonal vector converge towards the normal vector of a fitting plane. The next **Table 3-2** encompasses the results:

	Orthogonal vector (std)	**Orthogonal vector (mean)**	**Normal vector (fitting plane)**
X coordinate	(0.31, 0.01, 0.07)	(0.0906, 0.0005, 0.9405)	(0.0734, -0.0032, 0.9973)
Y coordinate	(0.21, 0.03, 0.32)	(0.0724, 0.0000, 1.0211)	(0.0643, 0.0003, 0.9979)

Table 3-2. Results of the test conducted on the linear distribution of the data (disparity in pixels, depth in millimeters, coordinate 'x, y' in pixels) given by a Kinect sensor. Notice that the numbers related in this table are rather negligible given the measurement units in which they are described. A global interpolation error introduced by this fitting plane over the data can be seen later in section Linearity in practice.

Method

Derived from the previous section, we aim here at aligning images from both sources. To do so, a spatial relation between coordinate systems will be set up. This relation in turn, is described by a 2D vector flow whose function basis (expected linear by far) needs to be calculated only once. After images have been aligned, color and depth can be merged into one 4-dimensional image [**22**]. Our method aimed at approximating this spatial relation using planar regressions is described as follows:

I. To sample as many planes as possible within the range of depth, several objects are placed at different distances in front of the cameras.

II. To capture nearly the same scene with the two cameras (color and range camera). Two images (I_c and I_d) are taken as synchronized as possible.

III. Sufficient landmarks are selected in I_c along with their peers in I_d. For each landmark threefold information is assessed:

   a) The x and y coordinates of the landmark in I_c, namely (x_c, y_c).
   b) The x and y coordinates of the landmark in I_d, namely (x_d, y_d).
   c) The depth of the landmark, namely D.

   Note that D is accessible likewise from I_d and pinpoints the distance plane on which the landmark was observed. Thus, $\Delta=(x_d, y_d)-(x_c, y_c)$ is but an example of the shifting of the images at distance D and not elsewhere. In general, each landmark provides evidence of the offset of the images at a specific distance. In practice, taking as many distinct landmarks as possible for a given distance D is advisable at all (as many distances as possible). As noted in the previous section the shifting Δ behaves linearly at D (i.e. disparity varies linearly into parallel planes).

IV. Now, the landmarks are used as a set of samples on which a global shifting function (Δ) can be interpolated. Eventually, this function can be regarded as a 2D vector flow describing the offset of the images. Hence, one function per coordinate is finally estimated and Δ can be reformulated as follows:

$$\Delta = (\Delta x(x_d, D), \Delta y(y_d, D))$$

Equation 3-5. Shifting function

The resulting function Δ is now vector-valued: it maps each point (x_d, y_d) in I_d to its shifted homolog (x_c, y_c) in I_c $\text{ShiftY}(y_d, D)$. So that, $x_d + \Delta x = x_c$ and $y_d + \Delta y = y_c$ for any given D. Yet, only few samples of this function are still known. Next section deals with the estimation of the model that best fits these samples. Also, this model will let us interpolate the function in its entirety.

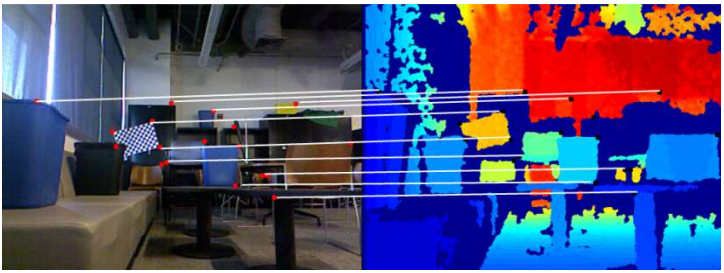


Figure 3-25. I_c (Left image) and I_d (right image). Red (I_c) and black (I_d) pairs of dots are landmarks manually selected. White lines coupling some of them make this figure more understandable. This amount of landmarks is quite enough for our method to work fairly well.

The Shifting basis function Δ

Notice that **Equation 3-5** has been conditioned to expresses Δ (shift) of a point in terms of D and its coordinates. Since the displacement of the images through the coordinates is known to be linear and the offset of the images varies linearly with depth too (for Kinect). Planes will be used to interpolate both, Δx and Δy. That said, the problem can reduce to a linear regression in a three dimensional space:

Let μ be either of the variables x, y; so that $\Delta\mu$ denotes either of the functions Δx, Δy. Given a set χ of n data points (landmarks) $(\mu_d^{(1)}, D^{(1)}, \Delta\mu^{(1)}), (\mu_d^{(2)}, D^{(2)}, \Delta\mu^{(2)}), \ldots, (\mu_d^{(n)}, D^{(n)}, \Delta\mu^{(n)})$. We want to find the equation of the plane that best fits our set. A vector version of the equation of this plane can be formulated as follows:

$$(\mu_d, D, \Delta\mu) \wedge a - b = 0,$$

where a is a normal vector of the plane and b is a vector that results from the product of a and the mean of the set of data points (i.e. $b = a \wedge \frac{1}{n}\sum_{i=1}^{n}(\mu_d^i, D^i, \Delta_\mu^i)$). Therefore, a turns out to be the only variable unknown.

Principal Components Analysis (PCA) can be used to calculate a linear regression that minimizes the perpendicular distances from the data to the fitted model [**181**]. In other words, given three data vectors μ_d, D and $\Delta\mu$, one can fit a plane that minimizes the perpendicular distances from each of the points ($\mu_d^{(i)}$, $D^{(i)}$, $\Delta\mu^{(i)}$) to the plane, $\forall i \in \{1, \ldots, n\}$. In short, the first two principal components of χ define the plane; the third is orthogonal to them, and defines the normal vector of the plane [**182**], namely a.

Alternative calculation using Thin-plate Splines

In the eventuality that the n data points (landmarks) do not show a linear distribution, we must use nonlinear fitting models to approximate Δ. As we will see throughout the next sections, in the practice this case is not rare at all. For cameras others than Kinect, approximation of **Equation 3-5** with a linear model is not very suitable (raw data is never linearized and operational depth range is larger). Also and more important, images distortion becomes a relevant issue in these cases. Cameras suffering from distortion are known to wrap the image with non-linear aspect [**162**]. Although in this work we use an adaptable class of splines [**183**], there is no constraint in this regard. The thin-plate smoothing spline f used in this work to approximate $\Delta\mu^j$ given a set of n data points or landmarks (μ_d^j, D^j), $\forall j \in \{1, \ldots, n\}$ can be regarded as a unique minimizer of the weighted sum:

$$\kappa E(f) + (1 - p)R(f)$$

Equation 3-6. General form of thin-plate spline

with $E(f) = \sum_j \left|\Delta\mu^j - f(\mu_d^j, D^j)\right|^2$ the error measure, and $R(f) = \int(|\partial_1 \partial_1 f|^2 + |\partial_2 \partial_2 f|^2)$ the roughness measure. Here, the integral is taken over all of $\mathbf{R}^2$, $|z|^2$ denotes the sum of squares of all the entries of z, and $\partial_i f$ denotes the partial derivative of f with respect to its ith argument. The smoothing parameter κ in **Equation 3-6** is derived from preprocessing of the set of data.

Let now f be the shifting function (also known as Δ in **Equation 3-5**), so that f maps $(x_d, y_d) \rightarrow (x_c, y_c)$ for a given D. The general equation for f is given as follows:

$$f(x_d, y_d) = a_1 + a_x x_d + a_y y_d + \sum_{i=1}^{n} w_i U(\left|(x_d^i, y_d^i) - (x_d, y_d)\right|),$$

Equation 3-7. spline-based mapping function

Here n is the number of samples (landmarks) we shall use to interpolate f, and a_1, a_2, a_3, w_i are the unknown coefficients we need to calculate. As for U, this is a special function underlying the thin-spline [**183**] defined as $U(x,y)=U(r)=r^2 log(r^2)$, with r being the distance

$\sqrt{x^2+y^2}$ from the Cartesian origin. Now, for the calculation of the unknown coefficients in Equation **3-7**we need to consider $r_{j,i} = \left|(x_d^i,\ y_d^i) - (x_d^j,\ y_d^j)\right|$, $\forall j$ and $\forall i \in \{1, \ldots, n\}$. Therefore:

$$K = \begin{bmatrix} 0 & U(r_{1,2}) & \cdots & U(r_{1,n}) \\ U(r_{2,1}) & 0 & \ldots & U(r_{2,n}) \\ \vdots & \ldots & \ddots & \vdots \\ U(r_{n,1}) & U(r_{n,2}) & \cdots & 0 \end{bmatrix}, n \times n;$$

$$P = \begin{bmatrix} 1 & x_d^1 & y_d^1 \\ 1 & x_d^2 & y_d^2 \\ \ldots & \ldots & \ldots \\ 1 & x_d^n & y_d^n \end{bmatrix}, n \times 3;\ \ V = \begin{bmatrix} x_c^1 & x_c^2 & \ldots & x_c^n \\ y_c^1 & y_c^2 & \ldots & y_c^{1n} \end{bmatrix}, 2 \times n,$$

and,

$$L = \begin{bmatrix} K & P \\ P^T & 0 \end{bmatrix},$$

where ^{T} is the matrix transpose operator and **0** is a 3×3 matrix of zeros. Following, let Y=(V | 000)^T be a vector of length n+3. Finally, define $W=(w_1, w_2,\ldots, w_n)$ and the coefficients a_1, a_x, a_y by the equation:

$$L^{-1}Y = (W \mid a_1\ a_x\ a_y)^T,$$

Equation 3-8. General solution of a thin-plate spline.

the solution of $L^{-1}Y$, gives all the necessary information to construct f .

Algorithmic performance, experiments and comparisons

In this section our algorithm for color-range calibration is outlined. Further, its computational performance is assessed too. Known issues relating to image distortion along with efficient solutions, are introduced and treated here in subsections (*Linearity in practice*) and (*Splines with no loss of efficiency*). Finally, comparisons with related methods are conducted in this section. It is worth noticing that though our method is proposed as a general framework to couple any depth-color camera pair, we have limited the comparisons in this section (*Evaluation of the method*) to a specific case where our algorithm may be specifically applied

as well i.e. internal Kinect calibration. The approaches whose efficiency is compared to that of our method in subsection (*Evaluation of the method*) are threefold:

- **A-1**: Calibration of kinect (Mapping of depth data onto the RGB images) using typical checkerboard-based stereo calibration [**184**] [**163**] i.e. assuming the range camera as digital.
- **A-2**: Calibration of kinect using the drivers provided by manufacturer (PrimeSense).
- **A-3:** Herrera's method [**174**] that uses a new depth distortion model to calibrate depth and color sensors.

3.4.3.1.1 Algorithm

i. Calculate PCA over the landmarks_x (firstly using x and D data only).
ii. Make a_x = third Principal Component, and, $b_x = a_x$ ^ mean (landmarks_x).
iii. Calculate PCA over the landmarks_y (Secondly using y and D data only).
iv. Make a_y = third Principal Component, and, $b_y = a_y$ ^ mean(landmarks_y).

v. Get I_c and I_d from corresponding sensors.
vi. Find delta_x using a_x and b_x in Equation **3-7** with $\mu=x$.
vii. Find delta_y using a_y and b_y in Equation **3-7** with $\mu=y$.
viii. Move each $I_d(x, y)$ toward $I_d(x+$ delta_$x, y+$ delta_$y)$.

Note that steps (*i*) to (*iv*) are performed offline and only once. Furthermore, if the system is ever decoupled, no recalculation of these steps is needed when recoupling. One can do the readjustment of the cameras by hand until acceptable matching of the images is attained. Further, we can see that the calculation of these offline steps is besides negligible. Typically, the computation of PCA requires eigenvalue decomposition (EVD) [**171**] using a Jacobi's method. Roughly, the overall PCA requires around $O(d^3 + d^2n)$ computations [**171**] (where n is the number vectors or landmarks and d is their dimension). Theoretically, in our method only 3 three-dimensional landmarks are needed (three points are enough to calculate a plane). In practice, however, the typical number of landmarks is approximately 20.

On the other hand, steps (*v*) to (*viii*) make up the whole workflow to be performed online. Particularly, we are concerned with steps (*vi*) to (*viii*) which are actually in the core of our computational approach. With a_x , b_x, a_y, b_y as constant data resulting from the offline phase, the solving of the linear model (steps (*vi*) and (*vii*)) requires little number of elemental operations [**185**]. Step (*viii*), in turn, is but a constant array assignation. Overall, the complexity of our online algorithm is linear with the size of the images N (i.e. O(N)). For images as I_c and I_d that usually don't exceed the order of Megabytes [**186**] the complexity is noticeably low.

Linearity in practice

Figure 3-26. Three mounted systems, from left to right: (RGB-Kinect) HDWebcam-KinectMicrosoft, (RGB-CamCube) HDWebcam-PMDCamCube, (RGB-SR4000) HDWebcam-SwissrangeSR4000.

Here below results of experiments conducted to endorse our model will be presented. Three camera systems were mounted as shown in Figure **3-26** (*RGB-Kinect*, *RGB-CamCube*, *RGB-SR4000*). Six different scenes (as described in section *Method*: I, II) were captured as follow: first four using *RGB-Kinect* and, the two remaining using *RGB-CamCube* and *RGB-SR4000* respectively. Following, corresponding shifting functions (Δx and Δy) were estimated for each scene. To do so, two fitting models were used and cross-validated (PCA-based linear regression and thin-plate splines). **Figure 3-27** and **Figure 3-28** show that in most of the cases both, Δx and Δy obey a linear distribution in the three-dimensional space *(Δμ, μd, D)*, with *d*={x | y}.

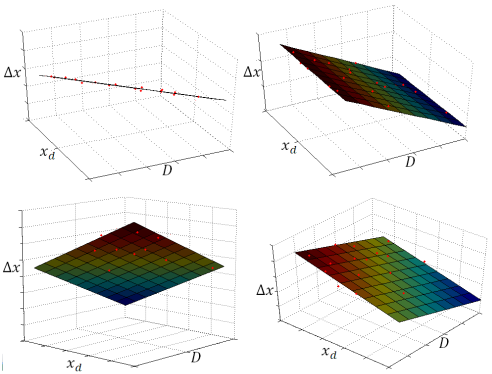


Figure 3-27. Six distinct scenes were imaged by three mounted systems. Scenes are represented by a depth map and a color image (Figure 3-25). Thus, each scene's landmarks provide two data sets to interpolate Δ_x and Δ_y respectively. A total of twelve data sets and twelve

interpolations were used. Firstly, twelve planes were used as interpolation models. Then, twelve splines were used in order to compare results. This figure shows the Δ_x and Δ_y data sets (red dots) as well as their interpolation models (surfaces), for two randomly selected scenes. All this interpolation models were cross-validated.

In some cases, the data of this experiment is fairly linear and well fitted by planes as expected (**Figure 3-27**). However, **Figure 3-28** that assesses the accuracy of the interpolations more objectively, reveals that this is not true for all the cases. First eight interpolations are accurate enough regardless the fitting model. Nonetheless, the last four interpolations were much better performed by splines. This means that non-planar surfaces do what planes cannot for keeping accurate fitting in these cases. Such a phenomenon occurs due to the distortion of the cameras which the last scenes were captured with (*RGB-CamCube*, *RGB-SR4000*). Further explanation on this regard is given in the next section. Also, to endorse the linearity in the first eight cases (*RGB-Kinect*), it is worth noticing that while both models fitted well. The splines presented a negligible roughness parameter ($\kappa \approx 0$, **Equation 3-6**). This means that they were almost flat (planes). This, of course, was not the case in the last four interpolations were κ was rather large.

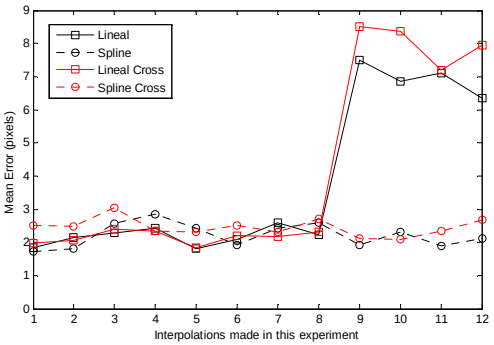


Figure 3-28. Validation: X axis represents the twelve interpolations made in this experiment. First eight interpolations belong to the scenes imaged by RGB-Kinect. The four remaining belong to RGB-CamCube, and, RGB-SR4000, respectively. Y axis represents the mean error of the interpolations using: linear models (planes), non-linear models (splines), 4-cross-validated linear models, and finally, 4-cross-validated non-linear models. Finally, the error marked by the crossed models is the mean error of the four validations.

Splines with no loss of efficiency

As shown in **Figure 3-28**, planar regressions failed to keep accuracy on fitting the last four data. However, these data are known to come from the scenes imaged by the SR4000 and CamCube cameras. Due to the distortion that these sensors present our linear approximation

is no longer suitable. At large, ToF cameras suffer from distortions both on the measured depth and on the ray direction [**162**] and Kinect is not an exception. This latter, however, is calibrated during manufacturing. The calibration parameters come internally stored and are used by the official drivers. This explains the fact that our method performs that well on processing Kinect data and not otherwise. While, this forces us to change the regression model in order to extend our method. We will show next, that this does not affect the computational performance at all.

To achieve the results shown in **Figure 3-28**, thin-plate smoothing splines [**183**] (described in section *Alternative calculation using Thin-plate Splines*) have been used to fit the surface underlying the data. The determination of the smoothing spline however involves heavily mathematical steps, such as the solution of linear systems. The solving thus usually takes a long time into our online routine (steps (*vi*) and (*vii*) of section *Algorithm*). In principal, this fact is drastically detrimental to our algorithm. To cope with this drawback, the regression model is no longer solved into the online-workflow. Instead, we sample (offline) the entire surface (be it a plane or spline) and store these values in memory. Therefore, the online process becomes independent of the mathematical model. Since, in any case, we no longer solve an equation but simply access the memory to read intended values. Eventually, steps (*vi*) and (*vii*) are lowered to elemental operations which enhance even more the performance of our method.

Evaluation of the method

Using Kinect, three patterns whose edges are known to be lines are imaged from multiple views. A set of 20 pairs (depth and color) of raw images is gathered in the end. A manual segmented version of all the color images serves as ground truth. Three areas are segmented from each image i.e. the three patterns. Moreover, we register the 20 pairs of images by shifting the depth maps using: A-1, A-2, A-3, and, the method being described in this paper. These shifted maps are automatically segmented in three areas as well. Finally, we compare common areas between these segmented maps (*sm*) and those of the ground truth (*gt*). Common areas must overlap exactly each other under the assumption of perfect registration. Thus, for each pair of overlapped areas (a_i^{gt}, a_i^{sm}) we assess its intersection ($a_i^{gt} \cap a_i^{sm}$), $\forall i \in \{1, 2, 3\}$.

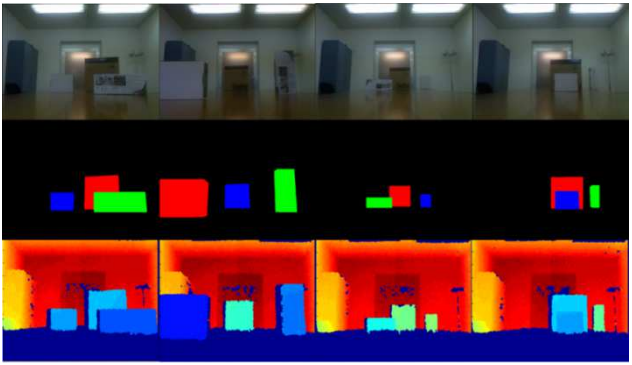


Figure 3-29. Some randomly selected images of this test. First row: Color images. Second row: manual segmented images (ground truth). Third row: Raw depth images.

An indicator of the accuracy (***Acc***) of certain method to register a depth-color pair of images is given by $\frac{1}{6}\sum_{i=1}^{4} |\frac{a_i^{gt} \cap a_i^{sm}}{a_i^{sm}} + \frac{a_i^{gt} \cap a_i^{sm}}{a_i^{gt}}|$. Notice that Acc is expected 1 for images successfully registered and below in other cases. We also measure the time elapsed during the registration of two images using the four methods. **Figure 3-30** and **Table 3-3** summarize the results of this section.

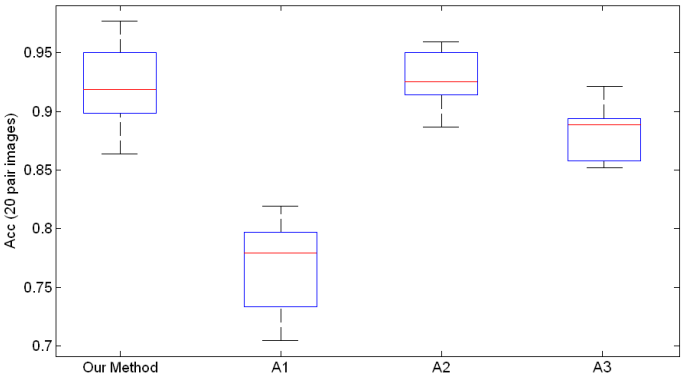


Figure 3-30. Calibration of internal depth-color camera pair of kinect (a specific case of registration) using three different methods. The accuracy Acc of our method for general registration of any depth-color camera pair is almost as accurate as that of the Kinect manufacturer.

An average *Acc* equals to 1 was not expected for any of the methods. This is mostly the case because segmented areas in range images are known to present highly noisy edges

(**Figure 3-29**, bottom row). Thus, flawless intersection with the areas in the ground truth (**Figure 3-29**, middle row) is unlikely. As a consequence, the accuracy of the manufacturer (A2) can be regarded as a baseline. By showing no substantial difference with this baseline, our method roughly reaches the maximum expectation of accuracy in this experiment. Moreover, having a standard deviation (STD) slightly smaller, results obtained with A2 may be regarded as a more consistent. Nonetheless, this very fact allows our method to achieve better accuracy than A2 in some cases (not outliers). As for A3, much better accuracy than A1 was noticeably reached, although the method certainly failed to surpass the threshold of 90% accuracy. This leads our method to a slightly better performance with nearly 92%. It is worth noticing that A1, A2 and A3 are methods that require extrinsic and intrinsic parameters of both cameras. Hence, use of extensive calibration techniques with checkboards and heavily manual work is unavoidable. The efficiency of our method suppresses these procedures, as well as maintains an average accuracy for otherwise unreachable.

	Our Method	**A1**	**A2**	**A3**
Time (sec)	0.021	0.036	0.027	0.033
Potential (fps)	46	27	37	30

Table 3-3. Computational performance. Our method is fairly efficient in computational terms. The potential fps number indicates the maximum rate at which the camera could process the images.

Finally, A2 is inextensible to the general problem of color and range images registration. This calibration is conducted during manufacturing and internally stored into the official drivers. Therefore, coupling the Kinect range sensor with an external color camera using A2, turns out to be of no use. On the other hand, A1 method does apply to the general problem. There is no apparent reason, however, to expect better accuracy by varying either of the cameras. The problem here relies on the treatment of noisy range-images (with not even visual features) as highly defined color-images. With regard to **Table 3-3**, it is worth stressing that both, our method and A1were implemented in Matlab, whereas, method A2 is an internal routine of the Kinect driver written is C++. Therefore, drastically better performance is expected in computational terms (CPU cost) for a binary compiled version of our algorithm.

3.5 Haptic-based Interfacing

Figure 3-31. Tactile-sound interfaces would allow unsighted individuals accessing information of an image via touch. The idea underlying here is that a user can hear elements of the real world by pointing (or touching) them with the fingers, as shown in this figure. Ideally, the user needs only to sweep (explore) the real world with his hands and fingers in order to get visual information. In an attempt to reproduce this idea, in See ColOr we capture the appearance of the real world into an image that is presented to the user through a tactile tablet. Thus, instead of naturally pointing into the real world, See ColOr's users need to carry a tablet and point (tap) into it in order to explore the sonified visual information. The implementation of this tablet-based interface will be discus in this section, whereas seminal ideas and early implementations of the ideal model (tablet-free) will be exhibited later in the conclusion section of this work.

In the local module See ColOr's resolution is 25 points (The sound of local and global modules). This small resolution gave rise to a drawback in See ColOr that is often likened to the tunneling vision phenomenon. Therefore, the global module allows the users reading the whole picture with their fingers by means of a tactile interface. In theory, the entire image resolution (460x640) is made accessible so as to make the most of the camera information. However, only as many points as fingers can be accessed simultaneously not to reach the limits of the audio bandwidth. **Figure 3-31** shows an ideal human machine interfacing which served as motivation to See ColOr's globule module: *Provided that contacted points (**Figure 3-31**) supply sufficient information (color, lightness, position, depth) coded into audio, one feels justified in saying that the unsighted user is getting into the visual world by means of his*

fingers. More specifically, he does so by means of the touch and audio trajectory playback [**129**].

3.5.1 Touch and audio trajectory playback

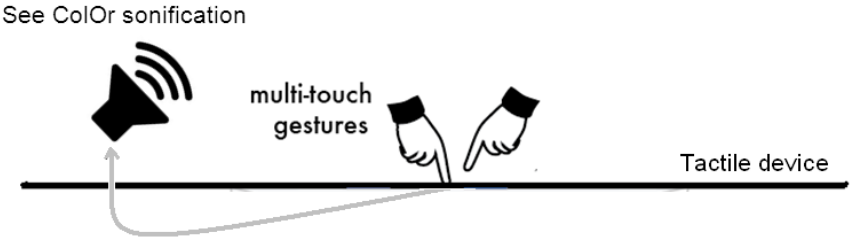


Figure 3-32. Touch and audio trajectory playback: navigating a touch-interface using the hand (finger) to move as a cursor, while triggering contextual information in the form of audio cues.

It is quite important to highlight that See ColOr exploits only audio feedback, given that its associated touch-interface provides no haptic feedback such as temperature, vibration, sense of texture and so forth (see **Figure 3-32**). Further than gaining haptic-feedback, therefore, here we use the kinesthesis [**106**] to promote proactive interaction of the user with the environment being explored (see **Figure 3-33**). After reviewing the state of the art of SSDs, one can learn that current prototypes all present unidirectional layouts (i.e. data flow exclusively from the system towards the user). By contrast, the inclusion of an interface grants a more proactive interaction to selectively explore, to discover points of interest, make comparisons, and, in general, enjoy a greater sense of independence. Overall, computational interfaces continue to be of great importance in HCI as they enlarge the legibility of the systems, increase the rate of information transfer and, allow the achievement of effective operation and control of the machine [**187**].

In that order of ideas we facilitate the interaction of a user with the nearby environment, with the focus on enhancing legibility, by providing an interface that meets two high level goals:

- Enhance environmental legibility by providing sound-encoded information along with interactive control to allow users to exploring the space dynamically. Eventually, mental representations of the environment layout may be built.

- Increase the transfer rate of information between the human and the machine, allowing more relevant content to be accessed at the same time (e.g. tapping with the

fingers on various points). This has to do with removing the tunneling vision phenomena in See ColOr.

Throughout this whole section, we describe and evaluate a muti-touch interface that allows the user to perceive color and depth of selected points, to discover points of interest and to develop strategies for exploration. Also, we address the optimal use of haptic and auditory trajectory playback concerning user interaction, and in particular the matter of the number of fingers needed for efficient exploration. Eventually, this interface may also increase the ease with which people can draw a mental image of the environment layout. Our general hypothesis is that the touch and auditory playback not only promotes proactive interaction but, it can be used as well in order to foster greater spatial awareness [**188**].

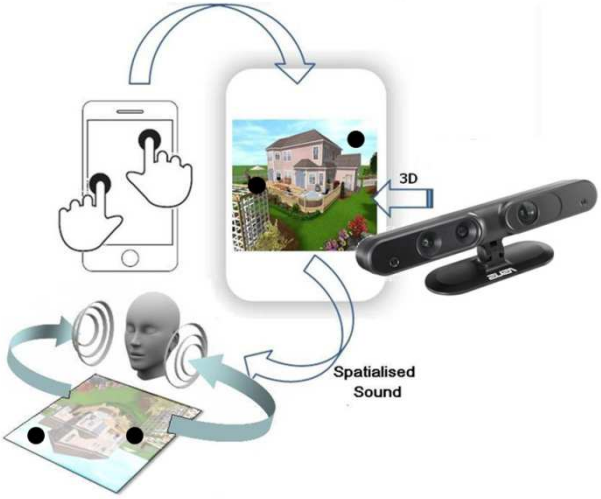


Figure 3-33. The interaction model that a touch interface aggregates to See ColOr.

Challenges in creating touch-interfaces for the blind

Touch-based interfaces are now present across a wide range of everyday technologies, including mobile devices, personal computers, and public kiosks [**189**]. In general terms, nonetheless, touch-based interfaces pose an especially daunting challenge as for the inclusion of the blind. While significant progresses have been reached in the accessibility domain to allow these individuals using mainstream computer applications. Touch-screens remain still inaccessible in many ways. Unfortunately, as discussed by McGookin *et al.* [**190**] the creation of accessible modifications and enhancements for touch-based devices is lagging behind the breakneck pace of mainstream development. Here, the principal drawback concerning interaction arises from the fact that in place of common devices such as the keyboard or mice,

screen computing generally offers a uniform, featureless surface [**129**]. This is perhaps the reason behind the exclusion of haptic interfaces in state-of-the-art SSDs. We argue, however, that instead of bypassing computational interfaces, the trend of research should address the rising of accessibility for the blind to gain better control of assistive technologies and in general, mainstream applications. Having no haptic feedback, an effective method to convey non-textual contents to the blind in a touch interface, is the use of sound.

haptic and auditory trajectory playback for spatial awareness

While haptic and auditory trajectory playback eases the interaction of blind users with touch-based interfaces devoid of haptic feedback; we believe that same multimodal strategy may be intended as a method to partially compensate visual cueing, when the interface represents a real world scene. Here, it is worth clarifying that haptic and auditory trajectory playback refers to the task of navigating a touch-interface using the hand (finger) to move as a cursor, while triggering contextual information in the form of audio cues [**129**] (**Figure 3-32**). Therefore, though no tactile cueing is received, the fingertip gives the subject a kinesthetic [**106**] understanding of spatial relations within the interface. These relations must be met likewise in the real world scene being mapped into the interface. In other words, emitted sounds represent and emphasize the color of visual entities in the environment, whereas finger's location reveals spatial relations thereof. As for the latter, this is especially appropriate to assess elevation given that See ColOr already uses spatialized sound to represent lateral positions.

3.5.2 A protocol for tangible interfaces (TUIO)

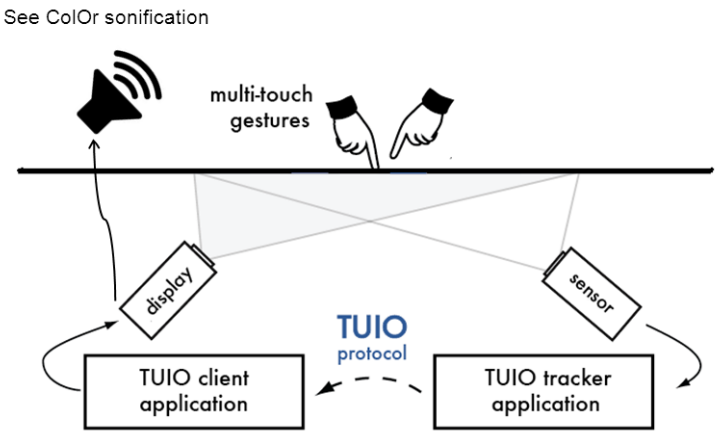


Figure 3-34. Handling model of a tactile hardware in See ColOr using the TUIO protocol

To create the haptic interface associated to See ColOr in this thesis we used TUIO [**191**]. Tuio is a simple yet multipurpose protocol intended to meet the necessities of tangible user interfaces. This protocol defines common properties to control finger and hand gestures performed by the user on the table surface. Tuio has been implemented within a fast and robust marker-based computer vision engine (i.e. reacTable [**192**]). Basically, this engine performs tracking and similar tasks including touch events and tangible object states. More importantly, this engine has been implemented on various standard platforms (supporting Java, C++, PureData, Max/MSP, SuperCollider and Flash [**191**]) and can be extended with multiple sensors. Reason why in See ColOr, the protocol has been used on iPad and earlier tactile tablet models. Tuio encodes data from this tracker engine (using a specific protocol) and sends it to a client application that is capable of decoding the protocol (**Figure 3-34**).

In short, the tracker reads into data (gestures) from the sensor of the screen, this data is then wrapped within a protocol and sent to a client that decodes the messages to generic interface events and draws the results into a graphical window in real time [**193**]. In See ColOr, however, displaying graphical data is not of much interest. By contrast, our Tuio-based client application discloses and outputs the data sent from the tracker, in form of spatialized sound (see **Figure 3-34**). Thus, in the end, visual impaired individuals that interact with a touch interface in See ColOr, receive no visual but audio feedback that they trigger with their fingertips (i.e. haptic and audio trajectory playback, **Figure 3-32**).

We developed two client applications in See ColOr: firstly, for the iPad, a java-based program was developed to receive the data (from the iPad tracker) wirelessly into the computer to be taken by a Matlab-based application for further processing (i.e. sonification). A jar file was made so that the java-program can be run directly through Matlab without the need of having a java-client running on the computer (e.g. Eclipse). As for the tablet Bamboo Fun Pen&Touch, modifications had to be made to the assembly-level driver to save the position of the fingers. Then, higher layer coded in C++ sends the data to Matlab for sonification. The whole TUIO-framework for developers is an open source project that can be downloaded from the web free of charge at [**194**].

3.5.3 Optimal interaction

Here we describe the results from a study looking at a two-hand interaction paradigm for tactile navigation for blind and visually impaired users [**195**]. To determine the actual significance of mono and multi-touch interaction onto the auditory feedback, a color matching memory game was implemented [**28**]. Sounds of this game were generated by touching a tablet with one or two fingers. A group of 20 blindfolded users was tasked to find color matches into an image grid represented on the tablet by listening to their associated color-sound representation. Our results show that for an easy task aiming at matching few objects, the use of two fingers is moderately more efficient than the use of one finger. Whereas, against

our intuition, this cannot be statistically confirmed in the case of similar tasks of increasing difficulty [**28**].

Inasmuch as the bandwidth of audio reaches their limits, it becomes imperative not to overwhelm the user with various emitted sounds simultaneously. In See ColOr this refers to the fact that the auditory pathway, even though useful for presenting some visual features through sounds, is severely limited when tasked with the analysis of multiple sound sources (i.e. representation of more robust visual information). This fact can dramatically impair the interaction of the user with a sound-based aid system. By contrast, people believe intuitively that multi-touch interfaces are better than mono-touch when interacting with tangible technologies, since more information of the screen can be accessed simultaneously. This intuition could succeed if we assume that the number of fingers is directly proportional to the rate of information being transferred [**28**]. Therefore, we attempt at assessing the objective difference on sound localization's ability of blindfolded individuals on a small tablet. We evaluate two cases: multi-touch mode (bi-manual) and mono-touch mode (one finger).

The game

In this game, pictures representing grids of colored squares must be explored. The task chosen for the study was for the user to find all the color matches. The pictures were made accessible through a multi-touch pad, as shown in **Figure 3-35**. The clues to lead the user to the goal were given by the audio cues being emitted from colored squares (touched with the fingertips). The actual location of fingertips on the tablet was also mapped into the spatialized sound. In other words, finger taps on the left of the tablet, produced sounds originating from the left, and likewise for the right side. Consequently, finger tapping around the center of the tablet produced sounds originating from the middle of the audio field.

Figure 3-35. A blindfolded individual playing the memory matching game of colors.

The difficulty of a this experimental game can be altered by using a set of grid images presenting 2x2, 3x3 and 4x4 squares. Thus, each grid contains a number of correct matches of 2, 4 and 8 respectively, with the 3x3 grid having one matchless square. A game meets its end when the participant successfully completes the three levels of difficulty. For each grid (level), the colors of the pairs were assigned after equal-spaced sampling of the Hue variable domain (360°) within the HSL cylindrical-coordinate representation [**29**]. Sampling the color space uniformly prevents from having repeated colored pairs into a same grid. Once a color was selected, its corresponding pair was assigned a random position into the grid.

Experiments

We conducted experiments on 20 blindfolded persons tasked to play the game as long as the three levels were completed. Recruited participant were given 10 minutes of training since they were not familiar with the color-sound code. In addition, four landmarks were attached to the touchpad indicating the middle of the edges of the sensitive area (see **Figure 3-36**). During the first part of the test, participants played the game through the three levels, using one finger. As for the second phase, the game was restarted to the first level with a new set of images in order to preclude transfer learning. This time the participant used two fingers to play, permitting the evaluation of multi-touch performance.

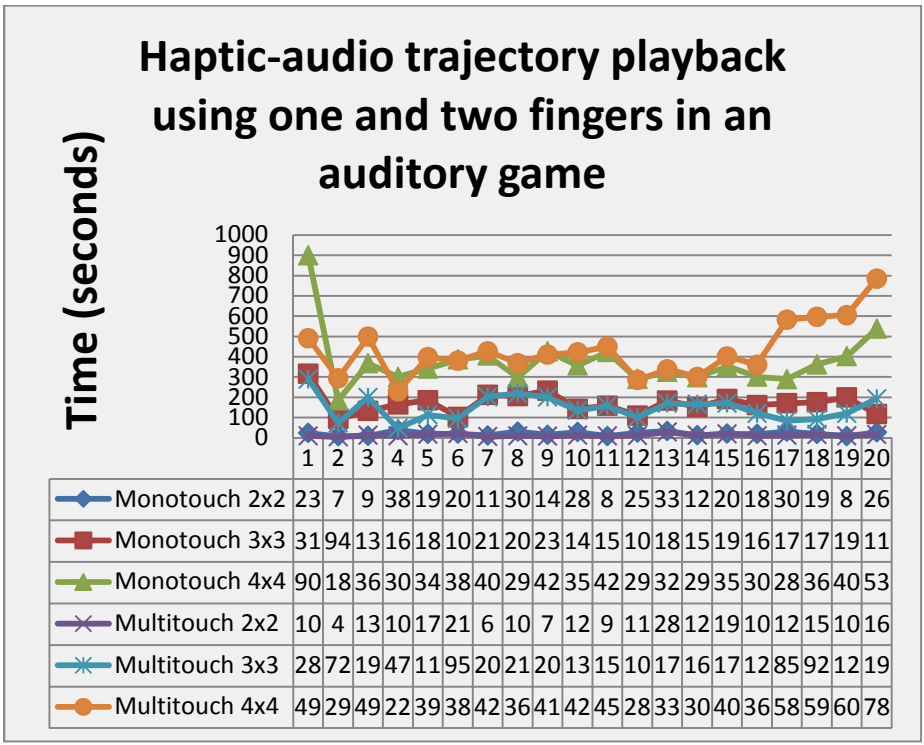
Figure 3-36. Touch pad on which grid images are represented during the game. The circles highlight the four markers indicating the middle of the edges of the sensitive area.

For each participant we measured the time required to succeed game levels (2x2, 3x3 and 4x4 grids) using both, one and two fingers. Some participants, though rarely, chose mismatched couples as well as pairs previously selected. Nevertheless, no relevance was given to this matter since memory capability was not a target of evaluation in this experiment. This is to say that a level was completed when the expected number of matches was reached, regardless mismatches or repetitions. Game participants were blindfolded for this experiment and

worn a high quality headset to properly perceive the spatialization of sounds [**196**]. They were not allowed to see the images before or during the test; a two minutes break was given between fulfilled levels, as well as between the mono and multi-touch sessions.

Results

Here, we report the results of the evaluation on matching ability based on haptic and audio trajectory playback using one and two fingers. A key aspect to assess the impact of mono and multi-touch strategies on the proposed game is the global time needed in both cases. Indeed, these data gives an insight into the advantages of one method with respect to the other, if any. The times spent by recruited participants while going through the three levels of the proposed game are shown in **Figure 3-37**.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Monotouch 2x2	23	7	9	38	19	20	11	30	14	28	8	25	33	12	20	18	30	19	8	26
Monotouch 3x3	31	94	13	16	18	10	21	20	23	14	15	10	18	15	19	16	17	17	19	11
Monotouch 4x4	90	18	36	30	34	38	40	29	42	35	42	29	32	29	35	30	28	36	40	53
Multitouch 2x2	10	4	13	10	17	21	6	10	7	12	9	11	28	12	19	10	12	15	10	16
Multitouch 3x3	28	72	19	47	11	95	20	21	20	13	15	10	17	16	17	12	85	92	12	19
Multitouch 4x4	49	29	49	22	39	38	42	36	41	42	45	28	33	30	40	36	58	59	60	78

Figure 3-37. Haptic-audio trajectory playback using one and two fingers in an auditory matching game.

For better understanding of the results reported in this experiment, we performed a paired t-test out of the data related to the times of interaction with one finger and two fingers. The paired t-test assesses whether the means of two series of experiments are statistically different from each other. By setting a hypothesis H_0 on the equality of the averages of the two data series, for the 2x2 game H_0 was rejected at a confidence level equal to 99%. Thus, for this level of difficulty the use of two fingers was significantly more efficient. As for the rest of the levels, the t-tests failed to reject H_0, therefore there is no significant advantage (statistically speaking) to use two fingers. **Table 3-4** shows the mean times (and standard deviations) needed for the participants to finish the games, whereas the results of our t-test are shown in **Table 3-5**

level	Mono-touch	Multi-touch
2x2	19.785 (10.024)	12.142 (6.261)
3x3	170.428 (58.949)	154.428 (65.493)
4x4	378.5 (163.295)	377.714 (80.315)

Table 3-4. Mean times (in seconds) to achieve the game's goal for each level, between parentheses the standard deviations.

level	t-test conclusion	p-value
2x2	Reject H_0	0.0087
3x3	Fail to reject H_0	0.1764
4x4	Fail to reject H_0	0.9821

Table 3-5. T-test results. Notice that only p-values < 0.01 can reject the herein-related equality hypothesis H_0.

After the experiment, participants were asked to give opinions about the differences experienced between performing the test using mono-touch and multi-touch strategies. Quite opposite to our intuition, 15 out of 20 participants described as irrelevant the use of either of the strategies: they invested pretty much the same effort to reach the goal in both experiment. Furthermore, the remaining 5 participants claimed to have felt uncomfortable with the use of two fingers when performing the last level of difficulty (4x4 grid). Relying upon this feedback, one could roughly say that inasmuch as the number of elements increases, the advantages of multi-touch strategy become unclear. This assumption does not match our intuition, since it was intuitively hypothesized that more fingers can access more information.

While there is a moderate gain when using two fingers at the first level of difficulty (2x2 grids). This is not reflected in higher levels of the game as confirmed by the t-test: no clear advantages in using two fingers have been found when the user is tasked to handle various elements of a touch pad.

In comparison with the human visual system, vision is a fugitive and dynamic phenomenon. The central area of the human retina has the best resolution for approximately two degrees. Since our eyes are very frequently moving to analyze our environment or a given picture, and by analogy if we consider that our eyes play the role of a single pointing device, it is worth wondering whether a pointing device such as a finger would be sufficient and necessary to mimic in a crude manner the human visual system. The results obtained during our experiments suggest that the improvement factor when using two fingers could be small or negligible for medium/difficult tasks. Is training the key parameter that will allow individuals to improve the time required to achieve a difficult task by means of two fingers? [**28**].

3.5.4 Building a scene in someone's mind

Here, the functionality of a Kinect sensor, accompanied by an iPad's tangible interface, is targeted to the benefit of the visually impaired by the construction of a detection/recognition system of pre-defined objects [**26**]. A features-classification framework for real time objects localization and for their audio description is introduced. Firstly, objects are extracted from the scene and recognized using feature descriptors and machine-learning. Secondly, the recognized objects are labeled by instruments sounds, whereas their position in 3D space is described by virtual space sources of sound. This enables blindfolded users to build a mental occupancy grid of the environment. As a result, they can hear explore and understand the scene using the haptic and audio trajectory playback from an iPad, on which a top-view of the scene has been mapped. Preliminary experiments using geometrical objects randomly placed on a table, reveal that haptic and audio trajectory playback can be used to build very accurately a scene in someone's mind in a satisfactory time, despite the absence of vision [**27**].

Scenery understanding relies on human vision mechanisms such as stereopsis, perspective unfolding, object identification and color perception amongst others. A fundamental research problem is the possibility of eliciting visual interpretation in the absence of vision, therefore by means of other sensory pathways. Along this line a number of researchers have been developing mobility aids to help visually handicapped users perceive their environment. Our hypothesis is that given a simple scene (in a controlled environment) composed of a limited variety of unknown objects with uncertain spatial configuration, a computer-assisted audio description can be used to build the scene in someone's mind so accurately that a physical reconstruction can be made [**26**]. The achievement of this audio description of the scene involves the encoding into sound data of information pertaining to both object identification and location. The implementation presented here linking a Kinect sensor, a laptop and a wireless iPad, will be generalized later in this thesis to experiments with unknown environments and unidentified obstacles.

Framework (using ortho-kinect)

A 3D Kinect sensor fixed at one extreme of a table (on which several elements lie) enables the machine-based object recognition. Afterwards, a top-view of the table is generated and mapped onto the iPad (the user experience is like perceiving the scene from above). Thus, the iPad emulates a miniature table and objects can be located within it, proportionally as in the real one (see **Figure 3-38**). Each object is associated to a specific sound so as to be distinguished from others. Sounds are triggered as the user touches the corresponding area on the iPad. In the end, the haptic and audio trajectory playback will help the user to build the mental map of the scene. In **Figure 3-39** a general framework to grasp the composition and layout of a scene using the haptic and audio trajectory playback is shown. The first stage that refers to **initialization (Figure 3-39)**, starts up all technical requirements such as, Kinect initialization [**197**], sounds database actualization, wireless communication between the iPad and the computer. Afterwards, the **input streams** (**Figure 3-39**) starts flowing and for every color-depth pair of images we perform the calibration process proposed in this thesis (see: Efficient registration of range and color images).

Figure 3-38. An illustrative example of the mapping of a real scene form a table into an iPad.

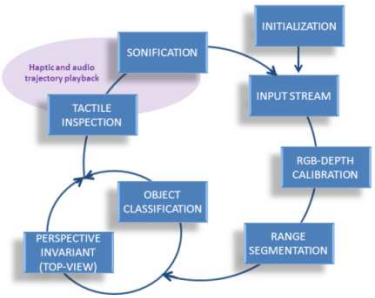


Figure 3-39. General framework for scene understanding using haptic and audio trajectory playback

Following, **range segmentation (Figure 3-39)** of the objects in the scene is carried out: Range images often enable fast and accurate object segmentation since objects are perceived as isolated surfaces with particular range. Thus, shape extraction is fairly good regardless color and illumination conditions. Here, the range of the Kinect sensor is partitioned into multiple layers. Afterwards, these layers are scanned one by one and surfaces (clusters) lying within them are labeled as objects. Farthest layers are ignored so as to filter out the background. The shape of a labeled object into the depth map can be then extracted. Color information thereof could be accessed by inspection of the same shape/area into the RGB map whenever a calibration has been previously done. This method attains precise segmentation in real-time for simple sceneries (i.e. without occlusions from the camera reference point of view). This process is illustrated in **Figure 3-40**, where a scene (bottom) is finally devoid of objects after segmentation.

The next step is the **object classification (Figure 3-39)**. After having segmented an object, this can be described by encoding its most representative features into scalar vectors. These feature vectors must be classified in order to identify objects with features alike, so that they very likely belong in the same class [**198**]. A wide gamut of vector descriptors can be found into the literature. Yet our descriptors are based on simple geometrical features of the objects, such as the perimeter, area, eccentricity, major/minor axis and the bounding box size. Given that this experiment uses sole geometric objects, more robust descriptors are unnecessary. Also, many machine-learning-based algorithms for data grouping meet the conditions to fit within this framework. However, feature vectors in this experiment were classified using a Multi-layer Artificial Neural Network [**198**]. Nevertheless, other recognition tasks (e.g. more complex objects) are expected to require more suitable methods for both, description and classification (this will be studied later in this thesis).

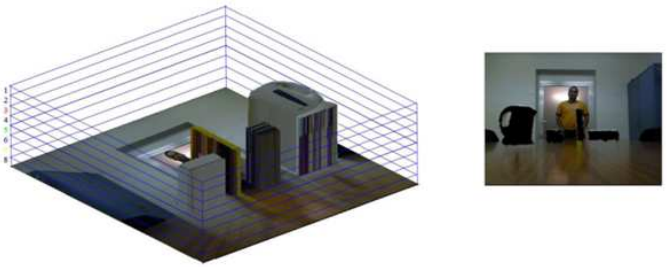


Figure 3-40. A depth-based segmentation process (objects are found lying within discretized layers of distance). We can segment and remove the objects (right) within the input image (left), while knowing the depth layer at which each of them was detected (1..8).

Finally, a key aspect in this experiment is the **perspective-invariant top view (Figure 3-39)** acquisition [**25**]. With the Kinect fixed at one extreme of the table, across progressive

ranges (distance, depth) the table's width seems increasingly diminished from the camera perspective (and the farther an object or the closer to the opposite extreme, the smaller it looks. See right image **Figure 3-40**). This is caused by the perspective effect (vanishing point [**25**]) and is normal in pinhole cameras. The result of the correction that needs to be done is to keep the table width fitting the image width at every plane (each plane of distance). Objects also must be stretched progressively and repositioned with the distance so that they preserve original size and location (see **Figure 3-41**). That's the equivalent of cancelling the effect introduced by the perspective. A top-view of the scene can be easily derived from an orthographic image (without perspective effect), given that we also known the depths (provided by the camera range i.e. Kinect). Next, we will review in more detail how the effect of the perspective was canceled in this work allowing the acquisition of such a top-view image.

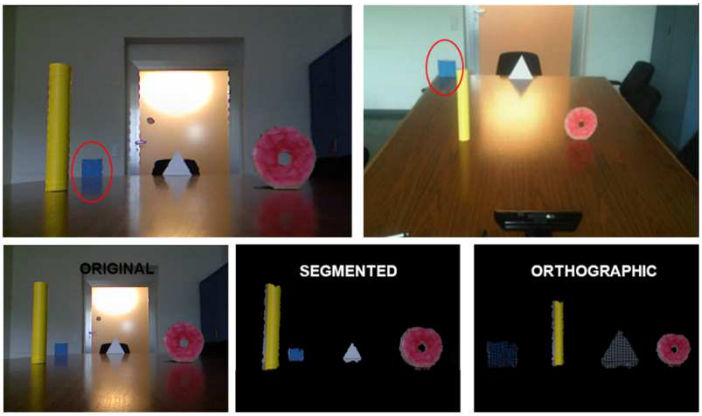


Figure 3-41. Perspective correction. Top-left image was taken using Kinect, because of the perspective the blue square (enclosed in red) seems to appear to the right with respect to the cylinder. Nonetheless, an image taken from above (top-right) reveals the actual location of the same square (i.e. to the left of the cylinder). At the bottom of this figure we can see from left to right: the original image again, a segmented version and finally, the orthographic version no longer affected by perspective effects. In this latter, we can observe the actual location (and size) of the objects without need of having an aerial view. In fact a synthetic aerial view may be built out of this orthographic image (see **Figure 3-43**).

Before building the final top-view image in this experiment, we need to achieve a perspective-free picture of the scene. This is doable if the rate by which the table decreases in size through the planes (with distance) is known, so that the correction of the perspective can be done by inversing this rate to keep the sizes fixed at needing. To do so, experimental data is necessary, in which pixel to centimeter ratio is to be found for each plane of depth. An object was used to measure the necessary experimental data [**25**]. That object was a box, it was put

in more than one layer of depth, and its size in pixels was measured. As the real size of the object is a known constant (15cm size of side), a table of pairs of values (the real size of the object versus its size in pixels) was made with 30 different values, 2 values per layer of depth, 15 different layers of depth was used. An example showing the process on 4 templates is shown in **Figure 3-42**. Then, the experimental data was interpolated to give the pixel-cm correspondence function (see **Equation 3-9**). This function is shown in **Figure 3-44**, whereas an illustration of perspective correction (aka orthographic camera [**25**]) can be seen in **Figure 3-43**.

$$y = f(x) = -1.4e-009x^3 + 8.8e-006x^2 - 0.019x + 17$$

Equation 3-9. Factor of conversion from pixels to cms (y) at given depth (x).This equation applies for the Kinect sensor and within its functional range.

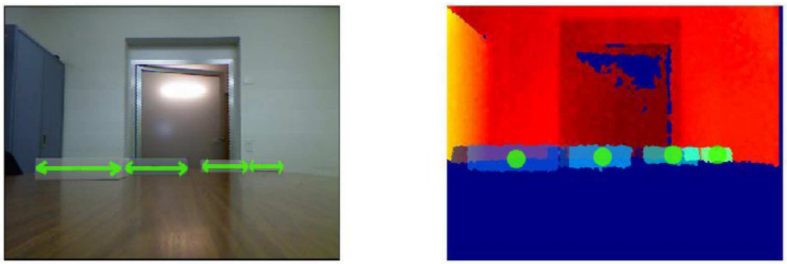


Figure 3-42. Measuring the perspective effect in a scene.

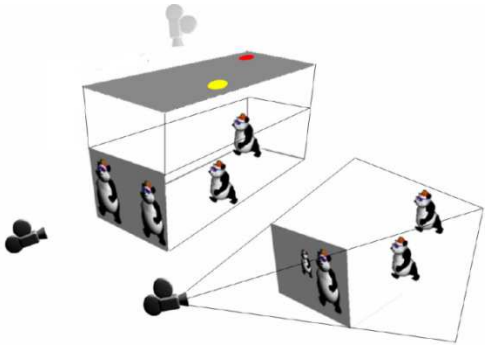


Figure 3-43. A perspective or pinhole camera (bottom-right): Objects further away appear smaller in size, besides the positions vary with the distance. An orthographic camera (top-left): Objects preserve natural proportions on size and position. Notice also that using this orthographic camera, a virtual ceiling mounted camera can be emulated. Thus, a top-view

image showing the location and distinction of the objects with colors (after automatic recognition) can be obtained.

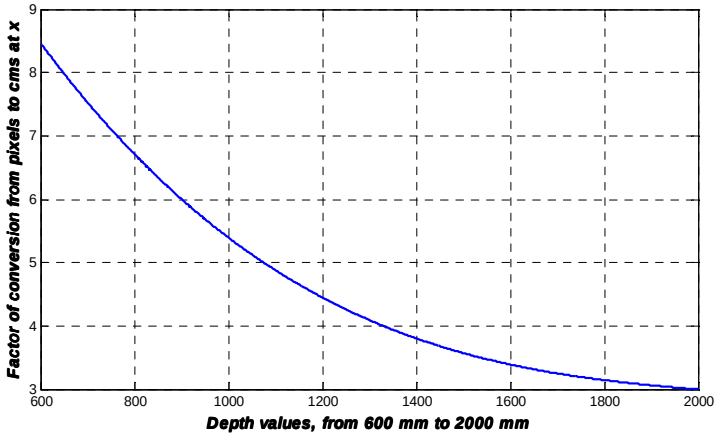


Figure 3-44. Perspective correction function for Kinect (Equation 3-9).

Last but not least, for the **sonification** (**Figure 3-39**) each class of objects is assigned a particular color/sound so as to be distinguished from others. Users must be previously trained to learn this object-sound association (audio icon). Then, objects recognized by the neural network are sonified when the user touches them on the iPad. While real objects' position on the table can be deduced by inspection on the iPad. Spatial virtual sources of sound used in this work create the illusion of sounds originating from the specific objects' locations in the real space [**104**]. It gives the user a more detailed idea of the scenery composition and the spatial relations between elements (see **Figure 3-46**). The end result of the whole process depicted as a general framework in **Figure 3-39**, can be seen in **Figure 3-45**.

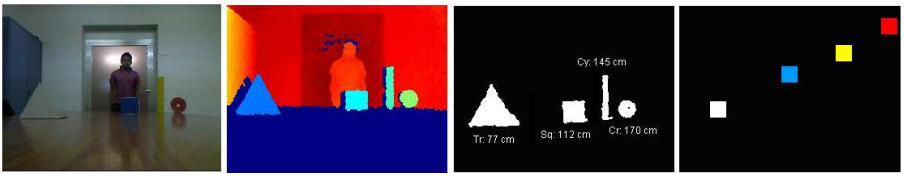


Figure 3-45. (leftmost) Color image (left) Depth image (right) Automated Recognition (rightmost) Top-View "multi-touch interface" with objects represented as colored squares. These colors will finally be associated with sounds.

Figure 3-46. Object location by haptic and audio trajectory playback. Besides, the user is given the illusion of objects emitting sound from the real position in space (right image). More precise location of an object can be achieved by haptic inspection (finger kinesthesia) on the iPad (left image). In other words, the user infers the location of the objects on the table out of his awareness of the fingers position within the tablet.

Experiments

Ten blindfolded participants were recruited to conduct experiments in two parts, for the first part participants had to recognize objects and locations using the iPad, see top-left image in **Figure 3-47**. As for the second part, the objects were removed from the table and the participants (without blindfold) attempted to put them back, see top-right image in **Figure 3-47**. Each participant was tasked to explore two scenes with three elements and two more with four elements. Training before the experiment was completely necessary as participants had no previous experience. At large, this training aims at making the participant familiar with the system and the color sonification. No more than 20 minutes were necessary in any case. The protocol used for the experiments is described as follow [**27**]:

Figure 3-47. Experiments on scene reconstruction using haptic and audio trajectory playback. Note that we also implemented this framework using the Bambo Touch Pad (bottom-left image).

- **Learning color-sound associations**

The participant is given up to five minutes to get acquainted with the color-sound encoding. In this experiment we only used four coloured objects so as to reduce complexity of this association process. Within 5 minutes of self-training, participants were able to distinguish objects based on their colors/sounds. To conclude this stage, participants were blindfolded and a quick test to ensure the success of the training was performed.

- **Tangible Localization Training**

To get acquainted with the tactile interface, participants were given another five minutes of training with the iPad. The goal of this training is to show her/him the sensibility of the touch screen and the precision needed to reach the elements within the interface. In addition, during this phase the user can develop strategies for haptic exploration. Typical strategies to scan the iPad are top-bottom zigzag, one finger fixed, spiral, top-bottom strips. The use of one or two fingers is unrestricted, whenever this should not produce any difference as explained in: Optimal interaction.

- **Auditive Localization Training**

For the last training stage, participants were given a practical understanding about sound spatialization. Several objects are placed on the table from the left to the right. Afterwards, she/he is allowed to hear the sounds originating from specific locations, accompanied by visual verification. The rhythm of the repetitions of the sound determining the element's nearness (i.e. the closer to the camera the faster), is also introduced through the same methodology.

This perhaps the easiest training, since representation of left/right object location using spatialization of sound is natural, and representation of depth by rhythm is quite intuitive.

- **Scene Exploration**

Recruited participants were blindfolded and the objects were randomly placed on the table. Then, the main part of the experiment took place; as an example see left image in **Figure 3-47**. Firstly, the participant had to find the location of the objects on the iPad by simple tactile inspection. Subsequently, she/he had to identify the objects nature out of the emitted sounds. The participant was expected to do a mental mapping from the iPad to the real world while exploring. Unless the participant claims to have achieved an accurate mental representation sooner, this exploration/identification process lasted ten minutes. In any case, right after the exploration was over, the elapsed time was registered.

- **Scene Reconstruction**

Objects were taken off of the table and then, participants removed the blindfolds. Following, participants were tasked to put the objects back on the table as trying to replicate the scene perceived during the exploration, see right image in **Figure 3-47**. This task was performed straightforward due to the mental idea of the scene gained during the exploration. Therefore, the time spent at this stage was negligible to be accounted. Also, no clue whatsoever was given to the participant who had freedom to reconstruct the scene.

Results

To allow an evaluation of these experiments, the top-view image of the scene is saved twice, before scene exploration (original) and after scene reconstruction (reconstructed) (see **Figure 3-48**). To assess the precision at which the imaginary scene was elicited in participants' minds, we compared those top-views. In the ideal case, they should match flawlessly as signifying perfect perception of the scene by the participant. In general, an objective estimation of the differences between these images can be achieved in this way. The Euclidean distance between objects within the first image (original) and their final location within the second (reconstructed) was used as a precision estimator of the reconstruction (**Figure 3-48**). Thus, the accuracy of the mental image elicited in mind can be objectively expressed.

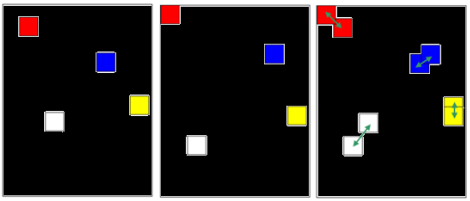


Figure 3-48. Evaluation method: Left image corresponds to the top-view of the original scene's layout. Middle image corresponds to the user guess. Finally, right image shows the mismatch between the two former.

Due to the calibration of the camera, the Euclidean distance between original and final object location in the pictures can be expressed in centimeters. Additionally in this experiment, this distance was normalized in order to express the performance in percentages according to the physical parameters of the experiment. To normalize distances, we divide them by the largest possible distance between two objects on the table which is the diagonal (260 cm). Hence, the separation of a relocated object with respect to its original position is expressed as a ratio to the largest mistaken possible separation. The results of the experiments are presented in **Figure 3-49**:

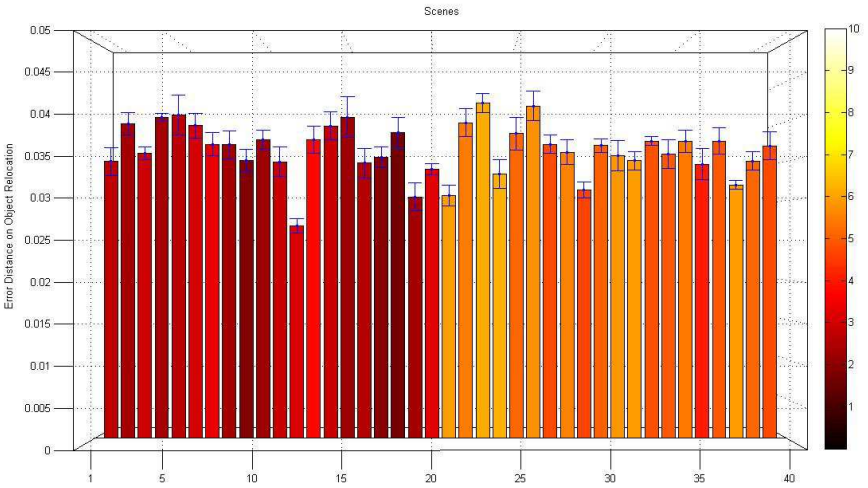


Figure 3-49. *X* axis represents 40 different scenes with three objects (1-20) and four objects (21-40). *Y* axis represents the average of the distances between the original and the final location of the objects. This average distance is already normalized. The colors of the bars (scenes) vary according to their exploration time that goes from 0 to 10 minutes (colormap). Each bar shows on top the standard deviation of its elements' relocation.

The results presented in **Figure 3-49** reveal that the participants were capable of grasping general spatial structure of the sonified environments and accurately estimate scene layouts. The mean error distance on objects relocation for all the experiments was 3.3% with respect to the diagonal of the table. This is around 8.5 cm of separation between an original object position and its relocation. In both cases (i.e. scenes with three and four objects) this

distance remained more or less invariant and was understood in this experiment as an estimator of accuracy. The exploration time instead, varied according the number of elements on the table. In average for a scene made up of three elements, 3.4 minutes were enough to build its layout in mind, whereas for scenes with four elements this time increase to 5.4 minutes. Such difference was given due to the increasing number of sound-colors associations to be learnt; the results showed no misclassifications of objects though. In general, all trial participants expressed enthusiasm as to the future of the project in this stage of See ColOr [**27**].

By and large, the results of this experiment make it feasible to extend scenery perception towards more general autonomous navigation aids. Thus, these experiments urged us to adjust the object recognition engine for general identification tasks (as it will be discussed later in this thesis). As to summarize, here we presented preliminary results in multiple object location and recognition through the use of haptic and audio trajectory playback. This experiment was meant to provide the visually impaired with a mental occupancy grid of the environment making use of a Microsoft's Kinect sensor, a laptop and a wireless iPad. For evaluation, the layout of a scene made up of a table and four geometrical objects was represented on an iPad and encoded into instruments sounds. This encoded information was accessed in real time using the fingers as stylus to trigger sounds (haptic and audio trajectory playback) [**27**]. The global information of the scene intended to be revealed to the user in this experiment is roughly summarized in **Figure 3-50**.

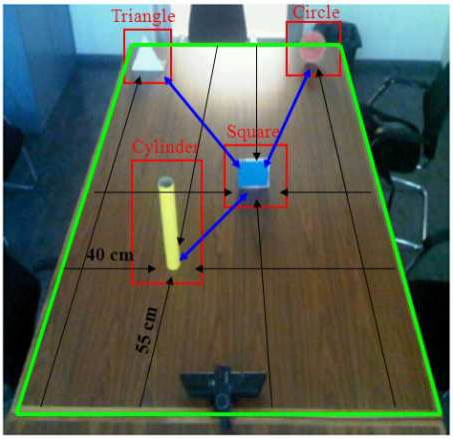


Figure 3-50. Information intended to be revealed in this experiment to the participants through the haptic and audio trajectory playback. This includes the class of the objects (red squares), spatial relation between them (blue arrows) and, spatial relation (black arrows) with respect to the table (green lines) and the camera.

3.5.5 Tactile Augmented Reality: An alternative to the use of a tablet

In Figure **3-51** we present a user who is interacting with See ColOr global module, which is based on a tactile interface hosted on an iPhone (or iPad). With the 3D camera on top of his head (helmet-mounted), the user is capturing a green plant that happens to be just in front. The image of the plant is then transferred to the iPhone screen, where it becomes accessible (or touchable) to the user with the fingers. The user in Figure 1 is triggering the sound of green, as long as his finger keeps on touching the leaves of the plant in the image. If the finger substantially moves for instance to the left, the triggered sound then will be that of white (the table color). Importantly, both sounds the green and the white, will be heard as coming roughly from the center and the left, respectively. For sounds in See ColOr are spatialized through the azimuth plane (i.e. left-to-right). Furthermore, the former sound will be emitted with higher rhythm because the plant is closer than the surface of the table behind (i.e. depth sonification). In contrast, if the user were to touch the frontal edge of the table, a higher rhythm will be assigned to the white sound. In short, the fingertip determines the point of the image that needs to be sonified. Therefore, the coordinates of the fingertip within the captured image is all we need to satisfy a user request.

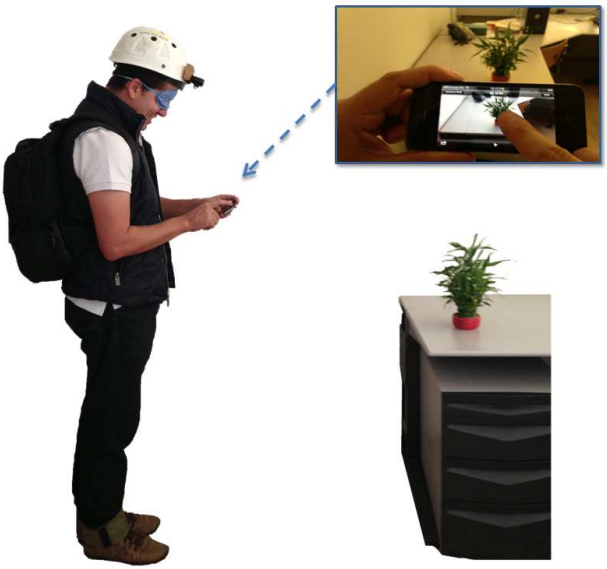


Figure 3-51. A user interacting with our global module based on a tactile interface. A zoom into his hands is shown top-right of this image to clearer see how the interaction takes place on the iPhone screen.

An issue related to the aforementioned tactile interaction is that the hands of the user remain occupied by the iPad, which dramatically reduces the freedom of experiencing the space physically. Likewise, the blind user has to rely on the screen limitations as an orientation guide to the environment. As a matter of fact, he needs to perform a rather complex mental mapping between the world and the screen coordinates to guess the real location of points in space. These problems have motivated us to find a more intuitive, simple and realistic way to interact with our global module. One idea arises from the fact already mentioned: all what is needed to sonify a point in the image are the coordinates of the fingertip within the image itself. Therefore, rather than using a tactile screen to sense the fingertip, we will have the fingertip itself show up in the picture. In other words, the user will be allowed entering his hands within the camera field of vision for us to track his fingertip. If the hand is outside or the finger fails to be detected, we switch automatically to the local module that sonifies constantly the central area of the image. To achieve a reliable tracking we will stick a marker on the user nail; this idea is well depicted in Figure **3-52**.

Figure 3-52. A user interacting with our handsfree version of the global module. On top-right of this figure, we display also the image captured by the head-mounted camera. Such image contains the user fingertip that has been enhanced with a marker.

Notice that in both Figure **3-51** and Figure **3-52**, the fingertip points the same spot within the Kinect-provided image (i.e. the leaves of the plant). The only difference is that in Figure **3-51** the fingertip reaches this point of interest through the iPhone, whereas in Figure **3-52** it does so by pointing in the real world. Additionally, in none of the cases the fingertip is actually touching the leaves, which adds one more similarity to these approaches. In principle for this example, the two methodologies exhibit only one negligible difference that is by no means detrimental to the interaction: using a tactile screen we sonify the point right beneath the fingertip, while in the tracking method we have to sonify the point just above the fingertip or marker. Though both points are expected to be quite similar, in any case, blind individuals cannot possibly notice such a small shift. Otherwise, there are indeed cases where the tracking method draws unquestionable advantages over the tactile one. Specifically, when the user points to a distant point not reachable with the finger (like the leaves), he simply gets as much information as though he was using the tablet: the instrument sound indicates the color, while the rhythm indicates the depth. However, in the case that the user actually touches the point in the real world (i.e. the point belongs to a reachable object). The user will get the natural sensations of touching (e.g. texture, temperature, resistance or elasticity) plus, the already mentioned color and depth. Thus, his tactile sensation will be augmented by color, a feature that has never been known to come from touch, but sight. We would like to introduce this concept in this thesis as Tactile Augmented Reality (see Figure **3-53**).

Figure 3-53. Tactile augmented reality. The touch of an object now produces a sound, the sound of its color.

Tracking the fingertip

In computer vision there exist a broad variety of strategies to track objects such as a finger. These may span from more complex like Kalman filters, to others simpler, yet efficient like convolutional patterns or color- and skin-based methods. To make it computationally inexpensive, here we will use a colored marker (or landmark) that wraps the finger so as to highlight it. Thus, we just need to perform some image processing to filter the bands of color we aim to detect and push away everything else. Likewise, some binary preprocessing and constrains on the sizes of detected colored areas will help. For instance, if we chose a strong pinkish marker (see Figure **3-52**), the color levels we need to filter at each frame of a video are shown in Figure **3-54**. These levels can be established in a test video from which we will manually sample the target color every time it shows up in the scene. More precisely, in Figure **3-54** we actually show two color space: the *RGB* (red, green, blue) which is the most used and the *Lab* (L is the lightness and *a* and *b* are the color-opponent dimensions) which might be more discriminative. As a matter of fact, we will see that color tracking based on *Lab* colors turns out to be far more reliable.

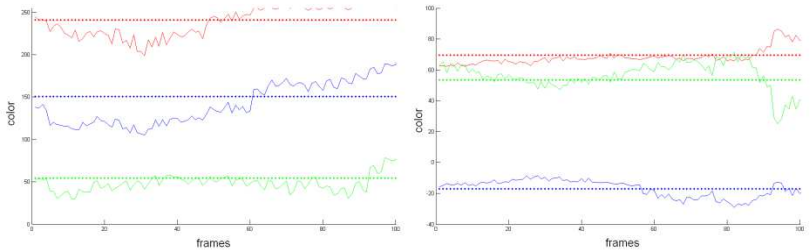


Figure 3-54. During a test video of 100 frames the pinkish color of the marker shown in Figure 2, had its color bands sampled each frame (manually). The curves represent the variation of each color band (or component) of the pinkish, frame by frame. To detect or segment pinkish areas in a new picture, each color component of the image must be filtered between the minimal and maximal variation reached by its corresponding curve in this figure. The dashed lines represent the mean value of each component during the 100 frames. Pinkish color is expected to be oscillating around these three mean values. To the left, we have RGB color representation, whereas to the right it is shown the Lab color representation.

In short, we will be segmenting the area of the image that presents similar tonality (pinkish) to that presented in Figure **3-54**. This is made possibly just by thresholding each of the color bands within their respective levels of interest (Figure **3-54**). Thus, each band will give us a thresholded area. In a binary image we represent the overlap of the three areas that made it through the filter or threshold. This overlapped area is but the pinkish marker, provided that no similar colored areas show up in the image. However, the area of the finger is expected to be of a certain size that could vary from: finger very close to the camera (seldom

seen) to very far from it (as far as the arm permits it). This adds a helpful condition to reduce false-positive detections due to colored background or artifacts. Once this area of interest has been segmented, some dilation and erosion pre-processing is applied to make it more compact. Finally, some elemental mathematical strategies must be applied to find the center and the top element of the area. This latter is always located on the area's boundary, right to the end of the major axis. We locate the point of sonification just few pixels above the top element of the area. Additionally, we can also use a two or three-colored marker, so that the process just described needs to be repeated three times. If the centers of the three areas are close enough, we would have found the marker. The likelihood of finding an area in the image that meets the same constrains in color, size and spatial distribution are just negligible.

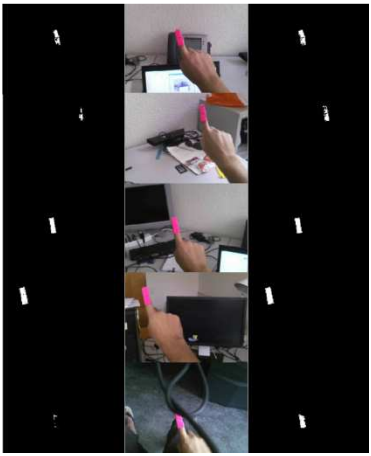


Figure 3-55. The segmentation of a pinkish marker stuck on a fingertip. The central column shows the original color images provided by the camera. The left column present the binary segmentations achieved over the RGB color space. In turn, the right column show the segmentations achieved over the Lab color space. Out of this example and experimental observations made in this work, it becomes clear that the segmentation on Lab color space turns out to be more robust.

Finally, it is worth saying that to sonify a point in See ColOr we need both its color and depth. We have seen how to track the fingertip that points the target of sonification within the color image from which, of course, the color will be extracted. Otherwise, to get the depth of the target point we just need to evaluate in depth map the same coordinates yielded by the color-based tracking. We expect a pixel-to-pixel correspondence between color and depth images, for we have previously used the calibration method described earlier in this thesis. Hence, images from both sensors are totally aligned, and no processing other than the color tracking itself is needed to get the depth of a pixel pointed by the fingertip. This concept is

summarized in Figure **3-56**, which concludes our idea of handsfree interaction with See ColOr global module:

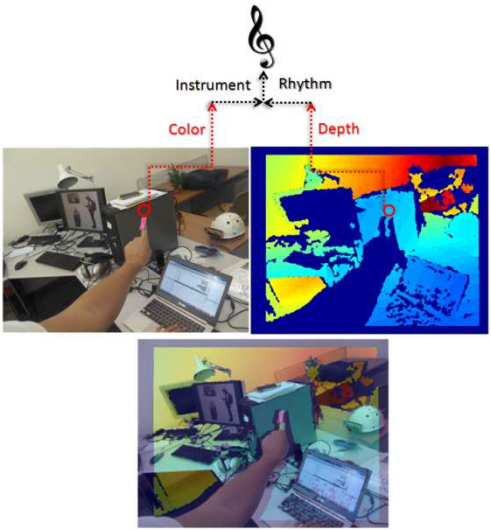


Figure 3-56. How the sound of a pixel emerges in See ColOr out of the natural pointing of a finger. At the bottom of this figure we show an image pair the kind of which we can obtain from Kinect (color and depth). Notice that this image pair is already aligned or calibrated, so a pixel-to-pixel correspondence is met. This being said, we can extract color and depth separately from both images. Color will be converted to an instrument sound by See ColOr, while depth will denote a rhythm of repetitions for that sound. Sound and rhythm emerge together as the sonification is produced by the pointed or touched spot. In the latter case, we will be achieving a tactile augmented experience for the user, who will know the color of the touched element.

3.6 Computer-vision-based visual substitution

3.6.1 Object recognition

To follow a target in a video stream either an object detector [**199**] or tracker [**200**] can be used. However, in principle, we have implemented a detecting-and-tracking hybrid [**201**] method for learning the appearance of natural objects in unconstrained video streams. Having a tracker and a detector running in parallel (during a learning phase) enables re-initialization of both techniques after failures. In other words, it is expected that the tracker corrects the detector (if needed) and vice versa. For instance, when the detections are weak or

the tracker drifts away (target temporally disappears). Thus, mutual information is used to build a more robust model of the target. Another key aspect of the use of a hybrid of this kind is that the learning is no longer based on a large hand-labeled training set. The target needs to be manually defined only once in a single frame. As a result, a first detector is built as well as the tracker gets initiated. Appearance changes of the target are learned online (frame by frame), allowing the detector to become more robust in time (with the accompaniment of the tracker).

That being so, both methods (tracker and detector) might continue to run in parallel while reporting the presence of the target. When it comes to blind subjects, however, this online approach introduces some difficulties. Firstly, the success of the method strongly relies on the first manual-made detection of the target. This fact dramatically diminishes the functionality, since the users would always need the help of a sighted individual to find an object. For that reason, a sort of long-term memory of learned objects is highly desired. Secondly, while this method works in real time, both tracker and detector are mono-target oriented. Hence, detection of more than one target simultaneously will overload the system, as new detectors and trackers would need to be included.

To cope with these drawbacks in See ColOr, we stop using the hybrid method once the detector has been constructed robust enough. Following, we save it in a database from which we can retrieve detectors upon request. When saved detectors are used, they are not accompanied of a tracker. Instead, we restrict the detection to the central part of the image so that the target is detected only when passing through this area. Besides allowing real time detection of multiple objects (several detectors running simultaneously in a small patch), this was done in order to give the user a spatial reference with respect to the target (if detected, the target must be right in front). This decision, however, gives raise to other issues reflected in prolonged searches (Experiments with blind individuals). We will address these problems in Improving time in experiments.

The tracker

We use a short-term tracker (that will become 'long-term' with the support of a detector) based on Kalman filter (KF) method [**202**]. First, a set of features points ($\mathbf{y}_{t-1}$) is sampled from a rectangular grid within the bounding box of the object (manually selected). Next, the KF tracks these points from one frame (t-1) to another (t). Based on a median over the new points ($\mathbf{y}_t$) resulting from the tracking, we estimated $\mathbf{x}_t$ the displacement and scale change of the bounding box. For each frame a new set of feature points is tracked, making the method very adaptive. The KF addresses the problem of estimating recursively the state of the variable $\mathbf{x}_t \in \mathbf{R}^n$ of a given continuous Markov process [**202**], from its observations $\mathbf{y}_{1,\,t} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$ obtained along the time. The process to be estimated is assumed to be governed by a linear stochastic difference equation [**200**]:

$$x_t = Ax_{t-1} + Vv_t$$

Equation 3-10. Observation model in Kalman-filter-based tracking methods.

with a $\mathbf{y}_t \in \mathbf{R}^m$ measurement y that is,

$$y_t = Hx_t + Ww_t$$

Equation 3-11. System model in Kalman-filter-based tracking methods.

Since mathematical models will fail to represent perfectly the evolution of a process. Besides, the process can be driven by disturbances that cannot be controlled in deterministic way. In **Equation 3-10** a stochastic term governed by a random variable $\mathbf{v}_t$ is also considered. This term denotes an independent Gaussian white noise sequence $N(0,\mathbf{Q})$. The matrix **Q** is regarded as the *process noise covariance* representing the inaccuracy of the deterministic dynamic model used. **Q** can be calculated as $\mathbf{VV}^T$, with **V** in **Equation 3-10** being the matrix that transforms the noise sequence to mimic the distribution of $N(0,\mathbf{Q})$. **A**, in **Equation 3-10**, is the matrix that established a linear auto-regressive relation between successive states of $\mathbf{x}_t$ (*system transition matrix*). **H**, in **Equation 3-11**, is the *measurement matrix*, and describes the deterministic linear relation between the state and its observations [**200**], [**202**]. Finally, the stochastic side of **Equation 3-11** signifies the disturbances that corrupt measurements and is determined by a random variable $\mathbf{w}_t$ (an independent Gaussian white noise sequence $N(0,\mathbf{R})$). The matrix $\mathbf{R} = \mathbf{WW}^T$ is known as the *measurement noise covariance*. At large, what the KF attempts to do can be regarded as the estimation of the conditional Bayesian probability density of $\mathbf{x}_t$ (i.e. $p(\mathbf{x}_t \mid \mathbf{y}_{1:t})$, which is necessarily Gaussian [**202**]). Thus, the conditionality here is given by the data $\mathbf{y}_{1:t}$. In short, the KF provides analytic expressions to compute the parameters of $p(\mathbf{x}_t \mid \mathbf{y}_{1:t})$ [**200**], [**202**].

Figure 3-57. Tracking of an object (a picture of Lena) using the Kalman Filter. The KF maps the sampled points (y_{t-1}) of an object (in one frame t-1) into y_t (the same points of the object in the frame t). It results in a sparse motion field (the dispersion of the points in the 2-dimension space). Based on this motion field, the scale and the displacement (x_t) of the object bounding box in the frame (t) can be calculated with respect to x_{t-1}.

The features

To describe points in an image (quite unlike standard approaches) we do not use SIFT [**100**] or SIFT-like descriptors, but 2bitBP [**199**] codes. These codes are short and fairly easy to calculate which allows online learning (with no training sets) and real-time applications. However, any other descriptor that meets these characteristics could be used as well (e.g. LBP 'local binary patterns' [**203**], haar-like features [**204**]). In fact 2bitBP are but a short version of these latter. A 2bitBP describes the area surrounding a point in the image by encoding its local gradient orientation (both vertical and horizontally). This description outputs just a 4 (2 bit) codes in contrast to 256 for standard LBP. Detectors based on 2bitBP achieve scale and rotation invariance inasmuch as they learn new aspects of the object online (frame by frame). **Figure 3-58** summarizes the calculation of a 2bitBP for a point in an image patch.

Figure 3-58. Here we present the four possible 2bitBP codes (a, b, c, d) that describe the surrounding area of a point in the image (around the eye). In 2bitBP approaches the scale an aspect ratio of this area are usually generated randomly (though rather small). Within this area four subareas can be extracted (A1, A2, A3, A4) when viewed horizontal and vertically. The final 2bitBP code is made out of the comparisons between vertical and horizontal subareas. The comparison is based on which subarea has a major integral value I (the sum of its pixels). Given that, in this particular example, I(A1)>I(A2) and I(A3)<(A4), the 2bitBP that describes the area is (c) (i.e. [01 11 00 01], with 00=gray, 11=white, and 00=black).

The Detector

We have used a Randomized forest algorithm [**205**] with the aim to decide about presence of an object in an image input. This algorithm belongs to the type of so-called real time detectors based on scanning window strategy [**206**]: the image is scanned across position and scales, at each scanned patch the presence or absence of the target is decided by a binary classifier (i.e. Randomized forest). Roughly, each image patch is described by a set of 2bitBP codes. The number of 2bitBP codes, as well as their size and aspect ratio are taken randomly. Also, these codes (or encoded features) are randomly partitioned into various same-sized groups. For each group, all its features are linked together in a feature vector x_i (binary) that points to a branch of a three ([**199**] , [**205**]) with posterior probability $\Pr(y=1 \mid x_i)$ [**205**]. Thus, a branch is represented by a posterior which, in turn, indicates the probability of a random event (e.g. the formation of the x_i pointing to it). The posterior is calculated by maximum likelihood estimator [**205**].

The patch is evaluated in the same way by several other threes simultaneously. Then, the posteriors or branches of all the threes are averaged, for the classifier to respond positively when the average of the probabilities is higher than 50% [**199**], [**205**]. Here, the posteriors represent the internal parameters of the classifier which are incrementally adjusted during the learning. Thus, each branch or posterior registers the number of positive and negative examples that have fallen into it, during the training. The randomized forest algorithm has been proved to have speed, accuracy and possibility of incremental update [**205**]. Figure summarizes the aforementioned process.

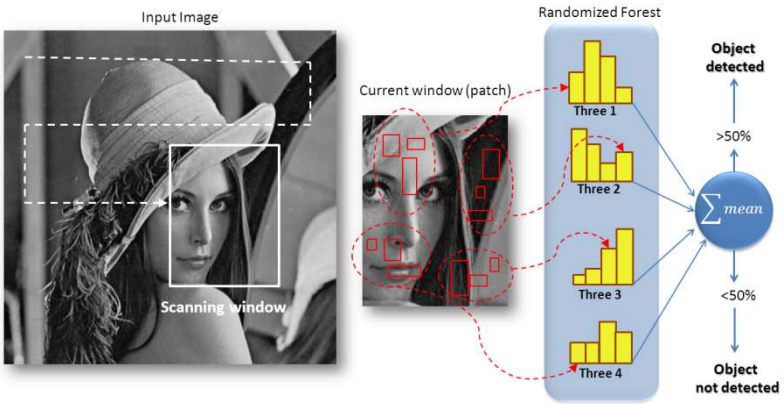


Figure 3-59. Here we show an example of a randomized forest classifier made up of four threes only. Each three has four branches, defined as posterior probabilities [205]. Inside the evaluated patch several points have been described using 2bitBP codes. The number of de-

scribed points, as well as the scale and aspect ratio of the 2bitBP are randomly selected at the beginning, though kept unchangeable. The coded points (or features) are partitioned into random groups of the same size. These groups are described in a vector x_i that points to the branch with posterior probability $\Pr(y=1 \mid x_i)$. If the average of the posteriors of the threes is higher than 50%, the object has been detected. No detection is reported otherwise.

The test

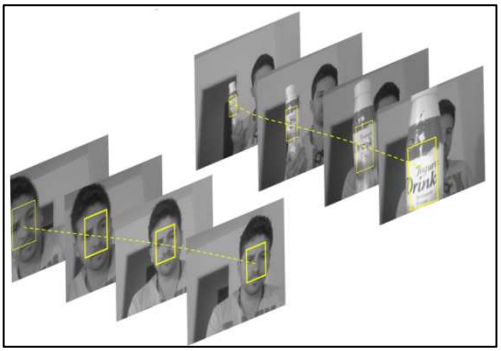


Figure 3-60. Two sequences exemplifying the performance of See ColOr's object recognition module.

Here, we present a study on natural object detection using See ColOr's recognition engine, which seeks to allow visually impaired users gaining awareness of certain objects they otherwise could fail to perceive, or simply need others help to do so. Notice that sometimes looking for an object (e.g. a fallen object) may end up in an embarrassing situation that might lower their feeling of dignity. In general, we attempt at letting an unsighted individual be aware of serendipitously encounters such as a person on his way to the toilet. Furthermore, we consider the case of conscious searches for daily objects such as a telephone, an exit, a trash can etc. Our recognition engine permits: (1) Learn the appearance of object during a learning phase that starts with the manually location of the object by a sighted individual. (2) When this learning reaches an end, the sighted person provides the object's name so as to label it. Next, the learned detector is then stored in a repository (database). (3) When the navigation/exploration is taking place, the unsighted user is notified about the presence of the object by the system which spells out the object's name every time following detection. To do so, the learned detector associated to the object is retrieved from the database and used upon user request.

In order to put this See ColOr's recognition module to the test, we surveyed five blind individuals asking about daily elements they regularly need and struggle to find. These persons were legally blind males (3) and females (2), meaning they have visual acuity of less than 20/400. Their ages range between 25 and 40 and all of them have educational level above high school. Participants engaged in this survey were recruited from the INCI (Colombian National Institute for the blind[12]). After this informative survey, we found three elements in common: A trash can, a telephone, and an exit way. For each of those elements, we (1) trained our engine to recognize it using an in-situ video as a *training sequence* of one minute length. (2) We shot in the same conditions a *testing sequence*. This was a four minutes video in which the object comes into view with significant alteration on appearance: rotation, perspective, partial occlusions, and scale. Also, we accounted for frame-cuts, fast camera movements and temporal disappearances of the object from the scene. Finally, we used our previously trained engine to detect/no-detect the object on this testing sequence. It is worth noticing that both, training and detection were performed in real time (simultaneously to the capturing of the *training* and *testing* sequences respectively). Some sample frames of the *testing sequences* can be appreciated in **Figure 3-61**. **Figure 3-62**, in turn, shows the precision-recall curves of our engine for each testing sequence of 3600 frames (15 fps x 4 minutes). We obtained these curves by comparing the resultant automatic detections (made online) with a manual labeling of the sequences (made offline).

Figure 3-61. Sample frames taken from real time detections videos (of three different objects) using See ColOr's recognition engine. The yellow square represents the area were the object

[12] http://www.inci.gov.co/

was detected. Rows 1 and 2: a trash can. Rows 3 and 4: an exit way. Rows 5 and 6: a telephone. These objects were selected as common targets by legally blind users.

The precision-recall curves (**Figure 3-62**) of our See ColOr engine for object recognition, show that the precision was kept between 72% (trash can) and 83% (exit way) at the total recall for the three cases. In addition, before the 50% of the total recall the precision in all cases ranged down from 100% till 90% only. While it is true that this system can recognize many objects that unsighted individuals might need in daily living, cognition extends beyond the field of object recognition. So stated, promising advances in computer vision as presented in [**207**] by J. Malik *at el.* should serve as reinforcement to our research on navigation aids. They use a poselet-based approach to attribute classification for describing people. In general, they are able to recognize gender, hair style and types of clothes in natural scenes (e.g., this person is male wearing glasses, jeans and t-shirt; he has long hair and no hat). In this way, quite a number of information that remains uncertain to the blind in everyday live could be cleared. Therefore, this would become more comfortable their daily live.

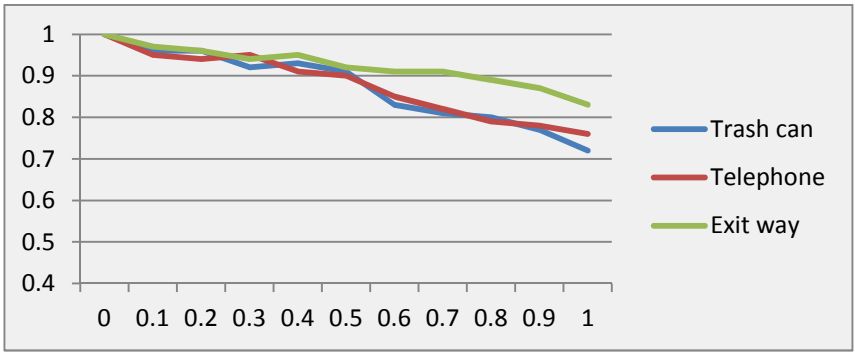


Figure 3-62. Precision-recall curves of See ColOr's recognition engine for three sequences of 3600 frames each.

3.6.2 Obstacles detection

The purpose of the alerting system in See ColOr purpose is to warn the visually impaired user when a threatening situation arises as consequence of an unexpected obstacle in his/her trajectory. Roughly, when a cluster of points in the video presenting a distance below 1 meter continues to approach over a given number of frames, the user must be alerted. Note also that the alerting system will run simultaneously with respect to haptic-based interface. Thus, users will keep on gaining context awareness rather than minding his step. As soon as the system launches a warning (alarm sound), the user is expected to suspend the navigation not

to bump into the obstacle. This allows the blind finding a safe, clear path to advance through (see **Figure 3-63**).

Figure 3-63. The alerting system. To the left, we can see a user passing through an unblocked door (clean of any obstacle). In the right image, instead, there is an obstacle (in this particular case the door itself) preventing the user from passing through. A collision is expected to occur if the user is not warned in time to avoid it.

The detection of objects lying on the user way (into which he or she is likely to bump) heavily relies on range image processing. At certain depth, within the range of the camera, we define a risky layer that limits the area that must be kept clear as the user advances. More precisely, this area (the risky area, **Figure 3-65**) extends from the user parallel plane up to the risky layer (plane) and in theory, none object should be detected within it. However, if an object were to appear within this area, it has to have passed through the risky layer before having entered the area (see **Figure 3-64**). This layer is fixed at 0.9 meters and it is constantly scanned to make sure it remains clean (nothing is entering the risky area). In other words, none entity must be detected (in the range image) within a depth of 90 cms as shown in **Figure 3-64**. Otherwise, presumably, there has to be something approaching the user (entering the risky area) and an alert should be launched. Consequently, this could be regarded as a primary alerting system (**Figure 3-64**).

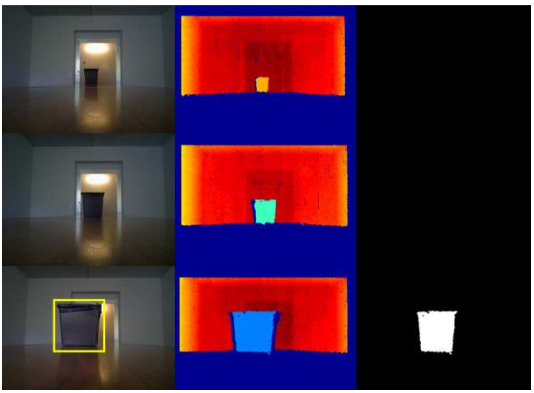


Figure 3-64. A risky layer. The first column shows three sequential images in which an object is approaching the user (the camera). The middle column presents the counterpart range images in which the depth of the scene is observed (depth is represented by colored layers). The rightmost column shows de binary representation of the range images using a threshold between 0.9 and 1 meter (i.e. 0.9<Range_Image<1.0). Therefore, these binary images reveal what is going on in the scene at the risky layer, allowing the detection of objects passing through it only. Importantly, objects in this layer (white clusters) are reported (yellow bounding box) only when they are large enough (e.g. more than 200 pixels). Finally, objects detected in this layer (and closer) have already reached a point so near the user that a collision is rather likely. Therefore, they might be considered as potential obstacles.

Importantly though, this entity that seemingly approaches might not be an obstacle, depending on the actual its actual trajectory. For instance, the frame of an opened door should not be taken as an obstacle (even though detected at 0.9 meters or closer), whenever the user is just passing through the door (**Figure 3-63**, left). In this case, a plausible conclusion is that the user will never bump into the frame, so that an alert turns out to be needless (**Figure 3-63**). In fact, we can apply the same reasoning to any object entering the risky area: even though the object is near enough, it might not be an obstacle if it is likely to pass just by the user's side. This idea is intended to prevent false-positive alerts and is graphically explored in **Figure 3-65**.

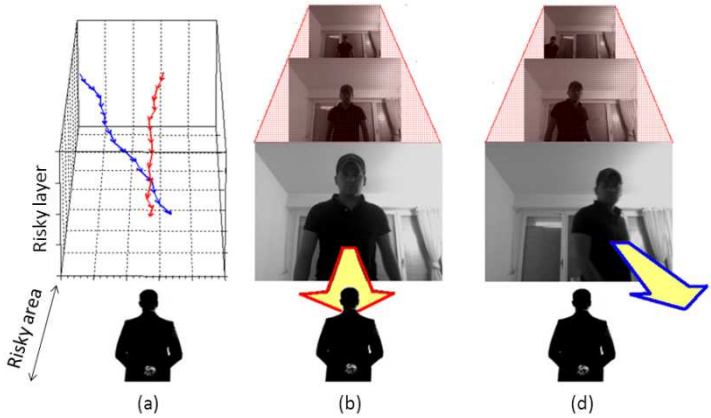


Figure 3-65. (a) The resultant path that two objects have traced after reaching the risky layer from deeper layers. Red vectors represent an object approaching directly over the user (situation (b)). By contrast, blue vectors belong to an object with diagonal trajectory (situation (d)). Also, (b) and (d) represent these two situations showing some sample images of the sequences that served to calculate the objects paths. In (b) a collision may be expected whereas, in (d), even though the object surpassed the risky layer, a collision is rather unlikely. Hence, only (b) should end up into a warning.

In our first approach once the object has been detected into the risky layer, we retrieve the last 15 color images back in time before the detection (note that in both, **Figure 3-64** and **Figure 3-65** only two of them are shown). Then, we use a short-term tracker (e.g. Kalman Filter [**202**]) to backtrack the object over this set of frames (see **Figure 3-65** (a)). Note that the initial target needed for the tracker initialization, is taken from the detection in the risky layer (bounding box shown at the bottom of the first column in **Figure 3-64**). Thus, we can trace the path of the object within the previous 15 frames (1 second before detected). Furthermore, using a vector average, we can estimate what its position will be when it reaches the user plane (1 second after detected).

A simple yet efficient second approach to the problem of obstacle detection is also explored in this thesis. We constrain the color video stream to the area that defines the spatial path he is walking through (area of interest). Therefore, we keep the sideway areas of the scene out of shoot and only objects standing on the user path (and not to the sides) are captured. For instance, the frame of an opened door would be excluded of the video when the user passes through (**Figure 3-63**), because the field of view of the camera accounts only for the area of interest. We do so by correcting the perspective of the camera as explained in detail in section Framework (using ortho-kinect) and in [**25**]. We simulate an orthographic camera that is no

longer affected by the perspective laws. This allows us to restrict the video into the area of interest (the one in front of the user) and keep the sideway parts of the scene out of shoot.

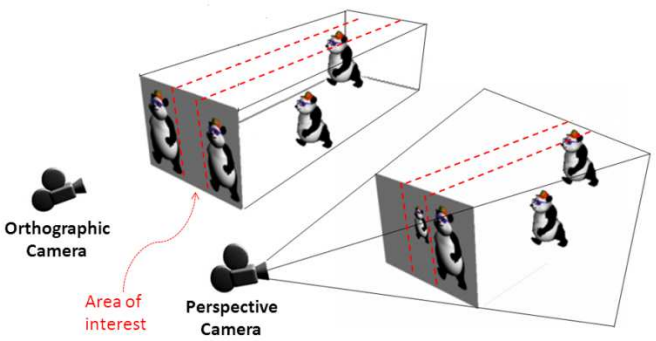


Figure 3-66. The central region of the video stream is defined as the area of interest when a user goes forward. Note, however, that in a perspective camera this area is 'contaminated' by objects that even though they are not actually inside, they seem to be just because of the perspective effect. In an orthographic camera this is no longer an issue.

While in the second approach we are able to limit the color video to the area of interest, we still need the range image processing to detect the obstacles that are seemingly approaching as explained earlier in this section (i.e. risky layer **Figure 3-64**). Therefore, we finally combine these two strategies (area of interest and risky layer) to build a robust alerting system. Importantly though, one can easily notice that when the analysis is limited to the risky layer and the area of interest, the orthographic view turns out redundant. None perspective effect can arise in one plane of depth (risky layer) because by definition the perspective appears across the multiple depths (progressively away). Therefore, an efficient method to alert the user about the presence of threatening obstacles could be as simple as the scanning of the central part of the risky layer. Note that doing so, the problem of the opened frame door (**Figure 3-63**) is solved since side areas of the risky layer are not scanned.

Finally, it is true that by using this strategy we lose track of the trajectory of detected objects. We confirm, however, that this fact does not affect the efficiency of the system: in practice, if an entity is less than 0.9 ms (risky layer) in front of the user (central area), it is convenient to alert the user (a collision is likely) even though the object is crossing diagonally (**Figure 3-65**). Therefore, report of false positives is not an issue in our system. **Figure 3-67** shows some sample images of a real-time video using See ColOr's alerting system to detect obstacles on the user (camera holder) way. **Figure 3-68**, in turn, shows some sample images of a video in which both, the alerting system and the object recognition were used. This combination will be tested in section Experiments. There are still some limitations in this alert-

ing system such as the manually calibration that has to be done to set the threshold on the number of pixels that form an obstacle. Also, the distance at which the object becomes dangerous (risky layer) has to be set manually. Sometimes, misdetections or false positives can occur too.

Figure 3-67. Two sequences in which the alerting system has detected an entity potentially leading to a stumble (or collision). First row, the users moves toward an unexpected obstacle (serendipitous encounter). Second row, the obstacle (a person) rushes over the user (threatening approaching). Once detected, the system segments the object and launches a warning message in either case.

Figure 3-68. A three-frame sequence of a video showing 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.

3.6.3 Reading text in See ColOr

During the last decades, screen readers have proved to be of great help to let the blind access textual information. Therefore, it makes all the sense thinking of 'world readers' to convey textual information (in natural environments) to the visually impaired. Usefulness of such technology that would make all this information accessible to blind is apparent. In everyday life, textual information represents one of the main medium for people to operate in their environment. Examples are numerous: finding the right shelf in a store when shopping, finding the right room in a hall, walking with precaution on a floor when a notice informs "Caution, wet floor", etc. Even if some of this textual information is translated into braille for the convenience of visually impaired people (floor levels in elevators, important information on medicine boxes), most of it remains inaccessible to them. This application could be also seen as kind of augmented reality: allowing a user to film a scene and getting audio information (in real-time) from the visual cues (text) contained on it (i.e. the written text is augmented with audio).

Our idea to implement a computer-vision-based module in See ColOr is extensible towards automatic detection, recognition and reading (by voice) of text. Fortunately, the research on text recognition from natural scene images has been growing recently. Many methods have been proposed based on a variety of image processing and image analysis techniques. Therefore, the interest of this thesis does not lie on the field of text-recognition research as such. Even though further progress still needs to be done (out of the scope of this work), nowadays there are sufficient available resources to enable See ColOr text reading. So, in this thesis we are rather interested in making use of state-of-the-art methods in this area to the benefit of the visually impaired. For that reason, here we aim at showing how available technology is compatible with our system to enrich the user experience and facilitate everyday life. Importantly though, we do offer a theoretical review to the subject below.

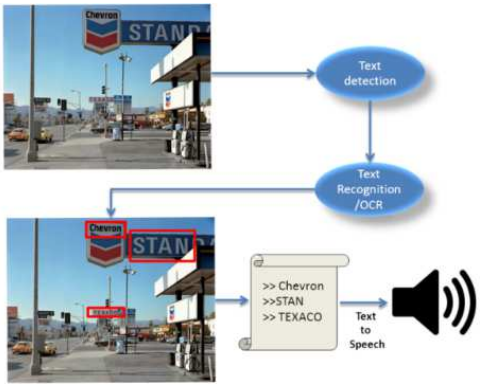


Figure 3-69. Here we present the general framework for text reading from natural scenes. Also, we added the "text to speech" process to focus the schema to the benefit of the blind. An input image is processed first by a text detector. Next, the text recognizer (usually based in machine learning methods) extracts the textual content (in plain-text) from the areas pointed by the detector. The recognizer could be an OCR (Optical Character Recognizer), though it is not very suitable for natural scenes. Finally, any text-to-speech software may be used to convert the textual information into voice.

Optical Character Recognition (OCR)

Optical Character Recognition is now a mature field of computer vision. It has been explored since the early days of digital computers: "the modern version of OCR appeared in the middle of the 1940's with the development of the digital computers. OCR machines have been commercially available since the middle of the 1950's [**208**]. Nowadays, anyone can have access to this technology. For example, the famous file hosting service Google Docs lets the user extract text parts of an image or PDF file [**209**]. We can also name a few open source OCRs, such as Tesseract, GOCR, and Ocrad. Also note that the software that comes with most of the recent scanners usually offer OCR technology. However, this prevalent technology only succeeds with highly constrained images, such as scanned documents, where the text is perfectly illuminated, contrasted, zoomed and aligned. In other contexts, its efficiency is drastically worsened. As a result, it cannot be used unchanged for the purpose of extracting textual information from natural scenes.

The need of methods more robust than standard OCRs is apparent. In other words an efficient text recognizer must succeed in recognizing text even with unconstrained images. There are two reasons why the encountered text could possibly not be well formatted. Firstly, more and more systems are deployed on mobile device, which means that the quality of the images might not be perfect. Indeed, images might not be perfectly illuminated and focused,

and the resolution might be poor. Secondly, the nature itself of a natural scene implies difficult illumination conditions (low contrast, reflections), as well as distortion caused by perspective, or rotated text. Also, the images can contain different text areas, with extreme size and font variation. In literature, this particular context is often mentioned as "scene text". Other terms employed are "unconstrained images", "urban scenes", or "natural images".

Text detection

Text detection, also called text localization, aims at determining the areas containing text in an image. Usually, the result of text detection consists in a set of rectangular windows surrounding the different words that have been found. They are called “text-boxes”, “text regions” or “text hypotheses”. The main issue when detecting text is avoiding false positive and false negatives as much as possible. False positives occur when the algorithm detects text in a region of the image where there is actually no text. Urban scenes can contain a lot of geometric shapes other than characters, which can lead algorithms to wrongly consider them as characters. Some papers, such as [**210**] with their F-HOG descriptor and [**211**]with their straight segment analysis, introduce methods that validate or invalidate text hypotheses. They help reducing the number of false positive. False positives can still be detected after the text detection, because the text recognition is likely to fail when applied over them. However, they are still going to be processed by the next steps, which can be resource demanding. As a consequence, false positives can increase the overall computing time. False negatives (or misses) represent a more problematic issue: they occur when the algorithm did not detect a text region of the image. They can be due to the properties of natural images explained earlier: low contrast, difficult light conditions, poor resolution, large size and font variation, etc.

Text recognition

Text recognition aims at extracting the text content of an image. The result of text recognition usually is plain-text. For example, text recognition applied on **Figure 3-69** could output something like: “CHEVRON STAN TEXACO”. This text can then be used by text to speech application, language translation applications (as in [**211**]), or it can be used to tag the image file with keywords in order to allow keyword search in a database of images (as in [**210**]). For instance, if we consider once again the application of assisting the visually impaired, text to speech would be the most interesting application. We could also imagine interfacing a system with a braille terminal.

Text detection and recognition: families of methods

Methods for detecting and recognizing text can be categorized in different ways. In this section, we review some of the families of text detection and recognition methods.

- Bottom-up / top-down: With regards to text detection, we can first distinguish bottum-up and top-down strategies [**210**]: “Bottom-up methods first attempt at de-

tecting individual characters and then merge neighboring positive detections. [...] Contrarily, top-down approaches directly look for text in images (sub) regions, mostly using a scanning window mechanism." The main difference between the two approaches is the computational complexity.

- Connected-component based / texture-analysis: Some methods first build connected-components from the images, i.e. sets of pixels that are all connected to each others. To build them, they proceed by searching for properties that are specific to text pixels. As mentioned in [**212**], these methods can be based on edges (because characters tend to have sharp edges) as in [**213**], greyscale or color homogeneity (because characters tend to have a uniform color) as in [**214**], or mathematical morphology (processing of the image based on the set theory) as in [**215**]. These methods are computationally efficient, but they run into difficulties when the text is noisy, degraded, multicolored, textured, or touching itself or other graphical objects, which often occurs in digital images [**216**]. Moreover, there are also some methods based on texture analysis. They consist in segmenting the image, in order to differentiate characters from background. To perform this segmentation, these methods compute information about the texture of the image (texture features), that help discriminating text from non-text. These features can be extracted "directly from the pixel's spatial relationships or from frequency data" [**217**].

- Standard OCR / End-to-end: Some papers only introduce text detection algorithms, and rely on standard OCR software for achieving text recognition. Contrarily, other papers introduce end-to-end methods. Most of the methods that are mentioned in this review belong into the first category. They proceed as follows: first, they localize text in the image. As noted previously, text localization usually results in a set of rectangular windows (text hypotheses). Then, they use standard OCR software for extracting the content of each text hypotheses. These methods have the advantage of being easy to understand and to implement. However, they are generally more computationally demanding, and less accurate. Instead, other methods introduce end-to-end algorithms: [**218**], [**219**], [**220**]. These methods are generally much more efficient, because the character recognition is designed to be robust.

General model of text recognition for See ColOr

On the one hand, it is worth recalling that text recognition requires good quality images, since letters may appear at reduced scales with the distance. Unfortunately, the quality of Kinect-provided color images is known to be rather deficient. These images suffer from low resolution, noise and unstable edges full of artifacts. Therefore, a key aspect to achieve reliable text recognition in See ColOr is the possibility to couple an external rgb-camera with the range sensor (Kinect). This process was described in detail throughout the section Efficient registration of range and color images. On the other hand, to recognize text in a scene, we

have used the open source implementation of the work "Text recognition in the wild" (TRitW), whose matlab-based code is available online at the website of the author Kai Wang[13] [**221**].

TRitW uses multi-scale character detection via sliding window classification based on a randomized forest algorithm [**205**] (similar to that used in this thesis for object recognition). As it was mentioned in section Object recognition this detector is quite efficient and allows real-time performance [**201**]. To train this detector synthetic data was generated by placing a small random character (with 1 of 40 different fonts) in the center of a 48x48 pixel patch and two neighboring characters, adding Gaussian noise and a random affine deformation (1000 images for each character, 62 characters). The efficiency of this approach to character recognition was evaluated using the ICDAR[14] Robust Reading Competition data set for real characters [**221**]. Note that here the text detector and the text recognizer as shown in **Figure 3-69** become one in TRitW (because the randomize forest method detects and classifies as well).

Additionally, this implementation (TRitW) also has a more sophisticated method for words detection based on Pictorial Structures [**222**]. Roughly, this method takes the locations and scores of detected characters as input and finds an optimal configuration of a particular word. This is possible by using a dynamic programming procedure over a tree data structure ([**221**] , [**222**]) built for a lexicon ([**221**] , [**222**]) (a repository of known expected words). The full implementation of this method (TRitW) for text recognition was tested in [**221**], using The Street View Text[15] (SVT) dataset (images and lexicon) that was harvested from Google Street View. Image text in this data exhibits high variability and often has low resolution. Some selected results on the Street View Text dataset can be seen in figure. Finally, it is very important to note that this TRitW is re-trainable or customizable depending on specific needs. In other words, one can create training characters and lexicons to train the method.

[13] http://vision.ucsd.edu/~kai/
[14] http://robustreading.opendfki.de/
[15] http://vision.ucsd.edu/~kai/svt/

Figure 3-70. Taken from [221]. Selected results on the Street View Text dataset. **TRitW** results are shown in green and words from the corresponding lexicons are shown in dashed pink.

To finally adapt the TRitW to See ColOr we used the TTS (text to speech) function in matlab that synthesizes speech from a string, and speaks it in mono audio format, 16 bit, 16k Hz by default. This function requires the Microsoft Win32 Speech API (SAPI). An issue that needs to be fixed in future adaptations for practical uses is the memory capability. Basically, running the text recognition method in parallel with the other See ColOr modules is not possible due to demanding computational processing. We tried to make it work simultaneously at least with the alerting system, yet we did not succeed either. This fact gives rise to the need of a switching functionality in See ColOr from exploration/navigation modes to text recognition only. While this is limiting because it prevents the user from unexpected text discoveries (while exploring), it is still useful for conscious searches (e.g. looking for an address).

At large, text recognition in See ColOr is feasible and promising; it requires the following line approach: an adapted external color camera bridges the user and the natural environment. However, there must be first an interface between the user and the camera (See ColOr as such). This interface provides three different modes of interaction for the user to select (local module, global module and now text reading). Note that when the user selects either the global or local module, both the recognition module and the alerting system will run in the background. By contrast, when the text reading mode is chosen, it will run independently (due to computational limitations). The text reading module closes the cycle of human-environment interaction by means of a text-to-speech app that gives audio feedback of the textual information in the environment. The detection and conversion of this textual information into speech is carried out by the 'text reading' module that internally has a structure as shown in **Figure 3-69**. This module can be eventually customized to particular needs. This will be done by semi-supervised learning using a data set gathered from the environment.

This data set is but a sample of the particular text we want to recognize given a particular situation. As it was mentioned before, in See ColOr we have adapted a state-of-the-art text recognition method already trained with natural street texts. **Figure 3-71** gives a theoretical representation to this approach whereas, **Figure 3-72** shows some pictures on this in practice.

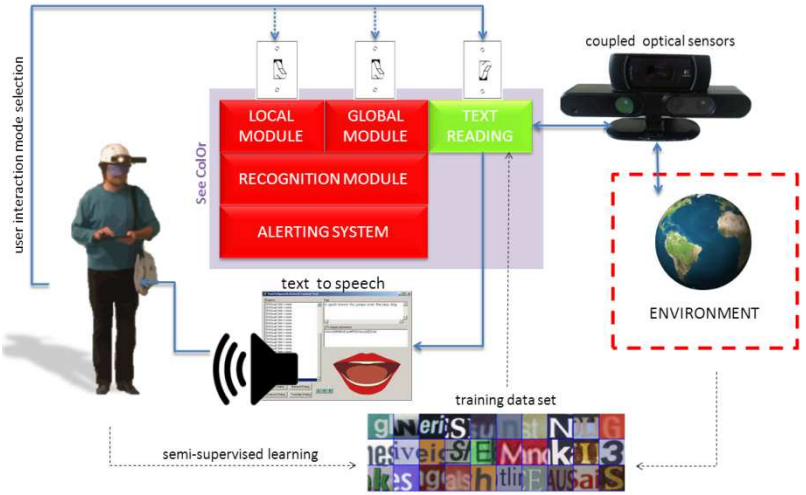


Figure 3-71. The addition of a 'text reading' module in See ColOr and the interaction between the user and the environment through this module.

Figure 3-72. For this experiment the **TRitW** was retrained using simple Calibri-font characters with white background. Also, we used a lexicon consisting in five words: {COLOR, READING, TEST, TEXT, SEE}. Finally, we use the function TTS (text-to-speech) in matlab to translate the results of **TRitW** into voice. In the left image there is a See ColOr user scanning the wall. Note that an external color camera has been coupled to the Kinect sensor.

Therefore, the user is enabled to switch from "text reading" mode (using high resolution camera) to exploration mode (with the range sensor) at any moment, and vice versa. In the right image we have extracted one frame from the video being captured by the user. We plotted on it the results of the text recognition at the time of the frame extraction (note the laptop in the user's backpack). These results are finally delivered to the user by synthetic voice. It is worth saying that even though the recognition is fairly good, sometimes it was even better (all the words were recognized)

3.6.4 Our approach to text recognition in the wild

In this section our attempt is to provide a framework to the problem of "text recognition in the wild"[16] [**222**]. Such a general framework will be adjusted within the context of our sensory substitution device See ColOr. More precisely, we would like to elaborate a general methodology to recognize text in unconstrained/natural scenes, supported by an implementation. Afterwards, we will adapt this approach to the needs of See ColOr (**Figure 3-71**) and finally, we will assess its efficiency through a particular example (i.e. a specific problem of text recognition). It is worth noticing, however, that both our general approach and its adaptation to See ColOr, will be unrestricted and fully adaptable to any other example. This means that they can be used to any text recognition problem just through re-training [**223**] (**Figure 3-71**). Formally, the goals to be achieved at this stage of the thesis are threefold:

- Implement a general framework for text recognition in the wild.
- Adapt such a framework to See ColOr.
- Apply the adapted framework to a particular problem.

Firstly, our implementation of a general text detector/recognizer will be based on well-known techniques such as deep learning and deep neural networks [**223**] [**224**] [**225**] [**226**] [**227**] [**228**]. Although, the manner we will make use of these machine learning methods is a completely original contribution of this thesis. As for the last goal above, the particular example we will chose to test the effectiveness of our approach refers to the problem (for blind individuals) of text reading at a bus station. Again, any other problem could be solved in the future with our framework due to its flexibility (i.e. re-trainability). In regard the aforementioned second goal, our general approach to "text recognition in the wild" will be adapted to the See ColOr's needs as follows.

Text recognition in See ColOr: Text recognition in the wild implies text detection and recognition (reading) in a natural/unconstrained image independently of both, location of the

[16] It is of vital importance not to mistake text recognition in the wild for standard OCR (optical character recognition) approaches. While the former applies to unconstrained natural images (mainly cluttered street views). The latter applies only to well-contrasted image documents with predefine characters type.

text and type of the text (font, size, view etc.). While this thesis is intended to fully fulfill the latter, in See ColOr the former has no clear advantages. For instance, if a blind person (seeking an exit) is told by See ColOr the word "Exit", for the system happened to find that word into the current scene, this person will gain no spatial awareness of the location of the actual exit. In sharp contrast, it will be far more useful if the seeker is provided also with a rough location of the exit such as right, left, front, bottom-left etc. In this view, we decided to use our multi touch interface (See ColOr global module) to let the user find the text himself with his fingers (Figure **3-73**). It is true that the location of the text could also be conveyed through left-right spatialization of the speech of the word (The sound of See ColOr). Nevertheless, our choice of a finger-based scanning strategy has also deeper implications to be discussed later:

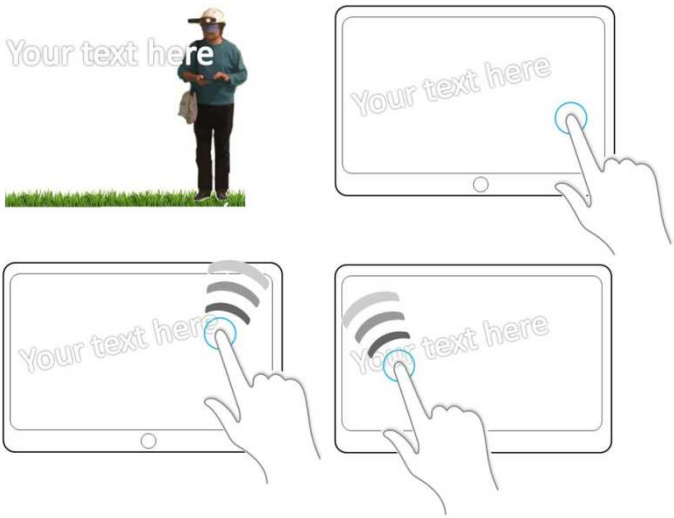


Figure 3-73. Text recognition in See ColOr. Our text recognition method constantly analyses the finger-tapped area of the image. If this area contains no text, no sound is emitted (top-right). In contrast, if the area does contain any text, the name of the letter being touched is spelled out with a text-to-speech converter (bottom-right and left). The text contained within the on-a-tablet-displayed image is captured by a head-mounted camera that records the natural environment (top-left).

To date, state of the art approaches in "text recognition in the wild" are far from being real time so that, given the scope of this thesis, changing this fact is none of our goals. Approaches to this problem fail to perform adequately in terms of time, mainly because current methods (including ours) are unable to scan a whole image rapidly enough due to computational limitations. Drastically improved results, however, can be achieved if the scanning is

limited to a small portion of the image (e.g. the portion cover by a fingertip). This being said, the aforementioned finger strategy turns up to be a good fit for See ColOr. Importantly also, this choice ends up in a real time application that can even be implemented in portable phones such as an iPhone.

In short, we will have a user scan the tactile tablet (e.g. iPad, iPhone etc.) on which the camera-provided image is being displayed, quite like the See ColOr global module. If it happens that the user touches a letter, See ColOr will immediately spell out the name of that letter making use of our text recognition method. This will allow the user to build words as he scans the content of the image, in quite similar way to the workings of Braille systems. Note that inside See Color's adaptation (i.e. without the finger scanning) our approach will succeed in any text recognition problem (though conditioned to prior training, **Figure 3-71**). Likewise, outside of See ColOr, in the general case of text recognition, our approach will succeed too. In this case, nevertheless, real time will be unachievable as it is for state of the art methods [**217**]. Our approach to text recognition in the wild, adapted to See ColOr, has been well portrayed in Figure **3-73**.

3.6.5 Deep Neural Networks and Deep Learning

Our strategy is rather simple; we will have two classifiers scan an image progressively through small areas or windows. This is broadly known as sliding window-based approach [145]. The first classifier (binary) decides whether the small area being analyzed contain text or not; i.e. text detector. Our second classifier then, decides which letter of the alphabet is contained in the area positively classified as text by the first classifier; i.e. text recognizer. The latter decision could be made by either one classifier of multiple classes (as many as letters), or multiple binary classifiers each associated to a particular letter; e.g. "a" or not "a". For the time being, let us adopt the strategy of having one classifier with multiple classes. In any case, however, the recognizer will be far more complex, so that having a binary text detector will prevent us from seeking letters in areas where not even text is likely.

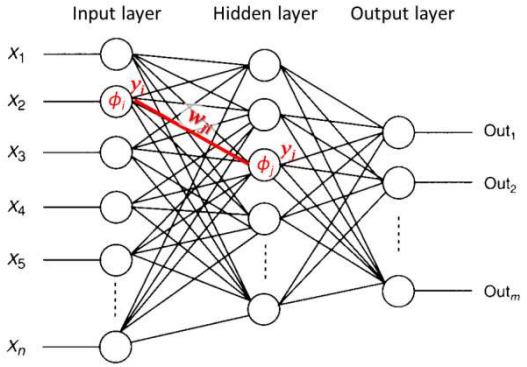


Figure 3-74. An Artificial Neural Network schema. We have highlighted in red a random neuron j within the network. This neuron is connected to all the neurons in the previous layer by weights w (e.g. neuron i is connected to neuron j through the weight w_{ji}). Each neuron in the network has and activation function ϕ (e.g. ϕ_j and ϕ_i). This activation function yields an output of the neuron known as a (or y in the figure); e.g. y_i and y_j. Notice that for simplicity, in this figure only one hidden layer was drawn. Nevertheless, there is no restriction in the number of hidden layers that can be used. As a matter of fact, in this work we will use networks with many hidden layers, also known as deep neural networks.

In this work both classifiers are Artificial Neural Networks (ANN) (Figure **3-74**) [**226**]. Artificial neural networks are models inspired by the central nervous systems (in particular the brain) that are said to be capable of machine learning and pattern recognition. By and large, they are represented as mathematical models of fully interconnected layers of neurons that can compute desired outputs from inputs by feeding information through the whole network (Figure **3-74**). Each neuron j is fully connected to each neuron i in the previous layer. These connections w_{ji} mimic dendrites and axons in natural systems. A neuron can be activated or not, when its connections (weights w_{ji}) multiplied by its inputs (or outputs y_i of the previous layer), are linearly pondered by an activation function ϕ_j. If the activation function surpasses a certain threshold, the neuron is then activated and its output (y_j or a_j) will affect the neurons in the next layer:

$$a_j = y_j = \phi_j(\sum_i w_{ji} y_i)$$

Roughly, ANN can be seen as a multivariable (weights) function that can approximate a set of desired outputs from a set of inputs so as to replicate the mathematical relation between them. This approximation is made possible by tuning its weights to specific values; i.e. training. Such training relies on a heuristic/iterative method, also known as the learning

method [**225**] [**226**]. A broadly accepted strategy for learning is the standard gradient descent [**227**]. This is typically the case, for the training of an ANN can be regarded as an optimization problem in which we want to minimize the error between the actual output of the net (fed with an input) and the expected output. This error (*E*) can be minimized only through modification (tuning) of the variables (weights) of the function (ANN). A well-accepted error function is the minimum mean squared error, defined within the context of ANNs as follows:

$$E(w) = \frac{1}{2}\sum_{l=1}^{u}\sum_{i=1}^{s}(t_i^l - a_i^l)^2$$

From this error function we must arrive (Appendix B, for details) to the general rule that updates the weights in an ANN:

$$w_{ji}(n+1) = w_{ji}(n) + \alpha y_i(n)\delta_j(n)$$

where $\phi_j^{'}(V_j(n))$ is the derivate of the activation function of the *j-th* neuron in the hidden layer evaluated in a local field. $\delta_k(n)$ is the local gradient of neurons in the output layer. And w_{kj} is the weight that links the *j-th* neuron in the hidden layer with the *k-th* neuron in the output layer. Note that the error is always being propagated backwards (previous layers). Therefore ANN trained with this rule are called back-propagation networks, since the tune the weights from output to input layer based on a gradient descend rule over an error function.

An ANN is called deep when it has many hidden layers (multi-layered), and it is called shallow otherwise. Training deep multi-layered neural networks is known to be hard. The standard learning strategy -consisting of randomly initializing the weights of the network and applying gradient descent using backpropagation- is known empirically to find poor solutions for networks with 3 or more hidden layers [**228**]. Nevertheless, complexity theory of circuits strongly suggests that deep architectures can be much more efficient (sometimes exponentially) than shallow architectures. Hence finding better learning algorithms for such deep networks could be beneficial. A clever strategy to train deep neural networks (known as **deep learning**) consists in simply initializing the weights (before applying gradient descent) not randomly but following a strategy. This strategy is rather simple: each layer but the last one, is trained individually as an unsupervised autoencoder. Once all the hidden layers have been trained, we apply standard supervised gradient based learning to the whole network. This will affect both the pre-trained hidden layers (fine-tuning) and also the untrained output layer. This pre-training strategy is known to improve on the traditional random initialization by providing "clues" to each intermediate layer about the kinds of representations that should be learnt, and thus initializing the supervised fine-tuning optimization in a region of parame-

ter space from which a better local minimum of the error function can be reached [**223**] [**224**] [**225**] [**226**] [**227**].

The unsupervised autoencoder [**228**]**:** An autoencoder neural network is an unsupervised learning algorithm that applies backpropagation, setting the target values to be equal to the inputs [**228**]. It is called unsupervised because the outputs need no labels as they are simply the same inputs, which is not a condition in the general case of ANN (Figure **3-75**). Importantly, the autoencoder is made up of an input layer, a hidden layer and an output layer. The former and the latter being of the same size as again, outputs are expected to be the inputs. The key aspect of the autoencoder has to do with the middle or hidden layer. In the general case, this layer is much smaller than the others two. This means that the network is forced to learn a compressed representation of the input; i.e. given only the vector of hidden unit activations (smaller than the input), it must try to reconstruct the original input in the output. Nevertheless, even when the number of hidden units is large (perhaps even greater than the number of input neurons), we can still discover interesting structure, by imposing other constraints on the network. This constraint is known as sparsity [**228**] (a variant to the gradient descent method), though given that in our implementation the hidden layer is always smaller, we will avoid this.

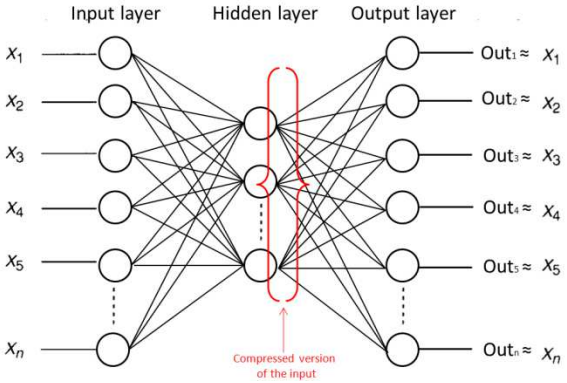


Figure 3-75. Autoencoder. Notice that the hidden layer must have a smaller number of neurons.

The main aspect here is that having compressed the input in the hidden layer, the autoencoder had to have learnt key features of the input that will allow its reconstruction. Thus, quite like PCA (Principal Component Analysis) [**228**], the autoencoder reduces a vector to its more relevant and characteristic dimensions; out of which the original vector may be inferred with minimum loss of information. This being said, we can go back to the idea of using autoencoders in deep learning to initiate the weights of a deep neural network. Each layer (*l*) of the deep network is to be treated as an individual autoencoder, by taking the previous layer

(*l-1*) as input layer and adding to it a temporary layer of the same size as output layer. After training, tuned (or trained) weights of this temporary layer are forgotten. In contrast, tuned weights between layers *l-1* and *l* are transferred, as the initial weights (between those layers), to the original network that is yet to be trained as a whole (fine-tuning). If we repeat this process layer by layer, we will be learning features of the input in the first layer, features of features in the second and so on. In other words, each layer is to encode a more abstract version of the input. The final layer of the original network that is meant to respond with the expected outputs of the global problem we are training for; will get tuned once the backpropagation algorithm is applied globally to the network. This routine is summarized in the following pseudo-code:

PSEUDOCODE FOR DEEP LEARNING IN DEEP ANNs

```
1. Initialize W randomly;
2. %pre-training
3. for i ∈ {1…L-1} do
4.     if i==1
5.         I_temp=I;
6.     else
7.         I_temp=logsig(W(i)* I_temp);
8.     end
9.     O_temp=I_temp;
10.    create a ANN_i of tree layers (size(I_temp), size(W(i)), size(O_temp));
11.    train ANN_i with I_temp and O_temp using standard backpropagation;
12.    W(i)=W_ANNi in hidden layer;
13. end
14. %fine tuning phase
15. train the original ANN with I and O using standard backpropagation;
```

W represents the weights of the original network to be trained as a classifier in this thesis. *L* is the total number of layers of the network. *I* and *O* represent the inputs and outputs respectively or patterns to be learned in the classification problem. *W(i)* are the weights of the layer *i* in the global network. I_{temp} and O_{temp} are the temporary patterns to be learned by auto-encoders-like hidden layer. Finally, note that a log-sigmoid function *logsig* has been chosen as the activation function for all the neurons in this network.

Reading at the bus station with Deep Neural Networks and Deep Learning

Here we would like to put to the test our approach based on neural classifiers in a practical situation of everyday life. In particular, one that turns out very useful for blind people who often struggle to use public transport in the city. We will go out there to Geneva's streets and collect real data to train our networks, and try to have them read text in real bus stops. Needless to say, that this is just a particular example we want to describe step-by-step for readers to be able to apply our framework to any other situation. Examples of the sort of images we can find in Geneva's bus stops are shown in Figure **3-76**. It is worth recalling here that text reading demands high quality images. Hence, we need to add an external color camera to Kinect, for our See ColOr prototype to acquire such images (Efficient Registration of depth and color images).

Figure 3-76. Sample images collected from some Geneva's bust stops. Our goal is to read with See ColOr any textual contents that may come across; e.g. Grange-Collomb, Tours-de-Carouge, Drize etc.

As it was mentioned before, we will have two networks analyze a given picture by small parts or equivalently, using a sliding window. The first network, will let us know which areas of the image do contain any text (text detector). The second one, in turn, will analyze those textual areas (using a sliding window as well) to tell the letters apart (text recognizer). Notice that since the classifier that tells letters apart is more complex than the simple binary network; having an initial detector saves a great deal of computational power. This idea has been portrayed in Figure **3-77**.

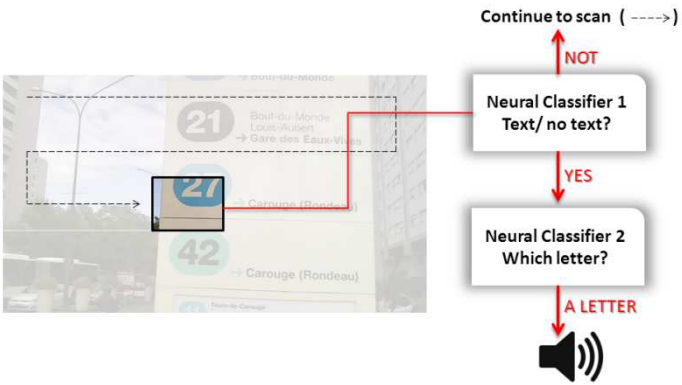


Figure 3-77. Sliding window strategy used in this work to detect and recognize text in the wild for a given image. The classifiers are deep neural networks trained with deep learning. Note that classifier 2 can be either a single classifier with multiple classes or various classifiers, one per class (letter). In any case it is far more complex than the first one which is simply binary. To save computational power, this recognizer operates only when we are certain that some text has been detected (first classifier's task). Finally, the two classifiers do not share at all the same window. If the first one detects text inside its windows, the second do explore the same area, though using its own windows yet to be described later on this work.

To train and feed our first network a 20x20 pixel window has been used. The content of this window is put in a 400 component column vector that serves as the actual input. Needless to say, these inputs need first to be normalized between 0 and 1. The output, in contrast, results from a 2-binary-neuron output layer that yields the decision: text (1,0) or not text (0,1). The overall structure of this network has been designed with 10 layers of 350, 300, 250, 200, 120, 60, 20, 10, 5 and 2 neurons respectively. Although, the choice for networks structures is known to be empirical and based on a trial-error process, we have a couple of reason for ours: pyramidal or size decreasing layers grantee that each layer pre-trained as an auto-encoder is indeed compressing the input; for the previous layer will always be bigger. This being said, the number of neurons per layer comes naturally, as we want to achieve realistic (therefore gentle) compressions not to force the network. Thus, we slightly decrease the size of a layer according to the previous one. Also this mild decrease in neurons each layer ends up in an actual deep network (i.e. many layers). As for the training, we use 3000 examples of windows containing no text and extracted from videos of three different bus stops in Geneva. Likewise, we use 2000 examples of windows that do contain the kind of text we aim to detect. These 5000 examples were selected and labeled (positive/negative) by hand. Actually, 10% of these data was used for validation and the rest 90% for training. Finally, for comparison we used the same data to train a shallow (not deep) neural network with standard backpropagation (not deep learning). In Figure **3-78**, 300 examples of each class of the training data are

presented, whereas training performance and precision-recall validation of both networks are shown in Figure **3-79**.

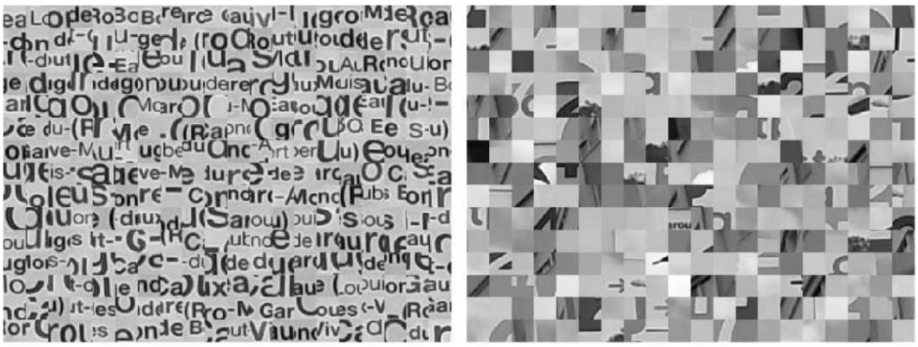
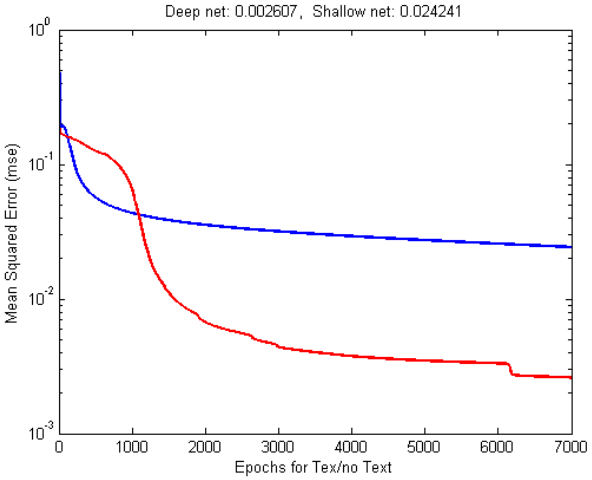


Figure 3-78. Classes: "text" or positive (left) and "not text" or negative (right). In this figure300 examples (20x20 pixels each) of each class are shown, the whole amount of positive and negative examples serves as a training data set for our binary network (detector).



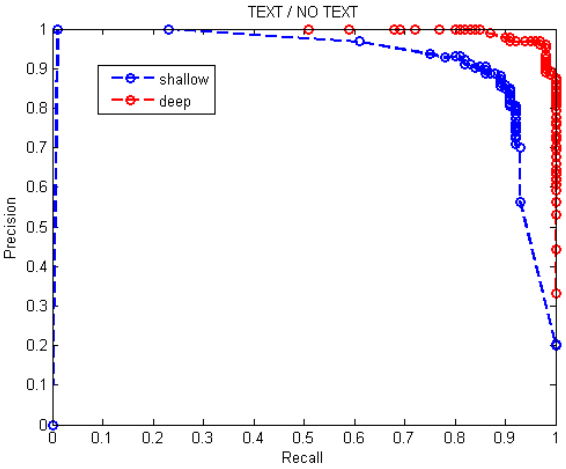


Figure 3-79. Top graphic represents the training performance of a shallow network of 350, 100 and 5 neurons (blue) and a deep network (red). Both networks were trained with the data presented in Figure 6 during 7000 epochs. Bottom graphic represents the precision-recall curves for the same networks in the validation test. The superior performance of the deep network is also reflected in the validation test. Both plots confirm that the deep network trained with deep learning is far more efficient (for this task) than a standard shallow network.

In the case of our second neural network the size of the window is not fixed, for we cannot possibly know the size in which a sought letter will show up. This size varies according to the distance, perspective and/or rotation that a letter may exhibit. Nevertheless, since an input of fixed size is required for this sort of networks, we normalize the size of the letters (whatever it is) to 16x13 pixels. This leaves us with a column input vector of 208 components or pixels. For training, this strategy works fine since example of letters are selected manually and regardless the size, they are simply taken down or up to the scale of 16x13 pixels. In sharp contrast, if the network is not being trained but tested, the selection needs to be done automatically. Since, again, we draw a total blank of the size of a letter, we need to try several different sizes of windows at each point the network is passing through. We will deal with this decision later on, though for the moment it is worth saying that this is one more reason why we need to have a text detector before the text recognizer (Figure **3-77**).

In this example we want our second network to learn just the most common letters, i.e. the vowels a, e, i, o, u (see Figure **3-80**). Thus, based on similar reasons than before, the structure of this network was chosen to be: 10 layers of 200, 170, 140, 100, 70, 60, 20, 15, 5,

and 2 or 5 respectively. Notice that the output layer may have either two neurons if we chose to have five networks that classify each vowel; or five neurons if we chose to have only one network that classifies the five classes of vowels. In this work will test both approaches: one multiclass classifier and several binary classifiers.

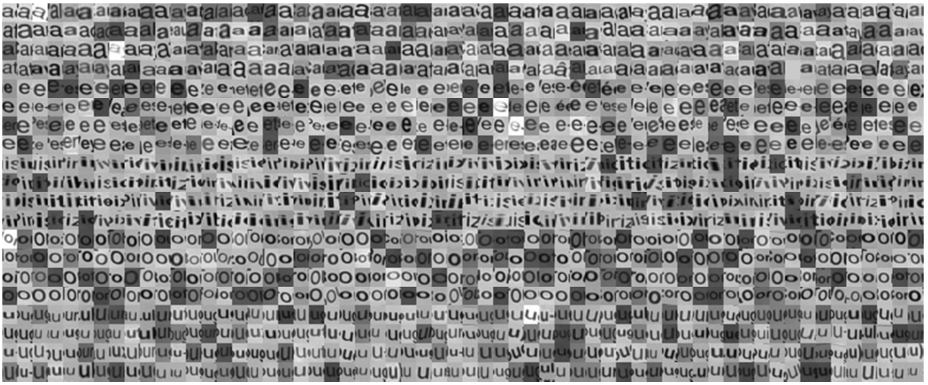


Figure 3-80. Examples of the five classes of vowels. Examples are normalized to a size of 16x13 pixels.

For each vowel we collected (out of the three video) 800 examples, which makes a total of 800x5 = 4000 patterns belonging to 5 different classes. In Figure **3-80** we show several examples of each class and 10% of these data was also used for validation. Notice that examples displayed in Figure **3-80** are already normalized in size (i.e. 16x13 pixels). If only one network is used to tell the letters apart, we simply label the 4000 patters with their respective classes and train the multiclass network. For otherwise, if 5 networks are used, we then train each using a leaving-one-class-out strategy [219]. This means that for instance, the network meant to classify the vowel *a* is trained with two classes one positive and one negative. This latter contains the rest of the classes in one: *e, i, o* and *u* (Figure **3-80**). Exactly as we did with the text/not-text network, we trained shallow structures for comparison. In Figure **3-81**, we present the training performance for five binary deep networks (one per vowel) and one multiclass (all vowels) deep network; as well as their shallow peers. Following, Figure **3-82** shows the precision-recall validation curves for networks with best performances in Figure **3-81**.

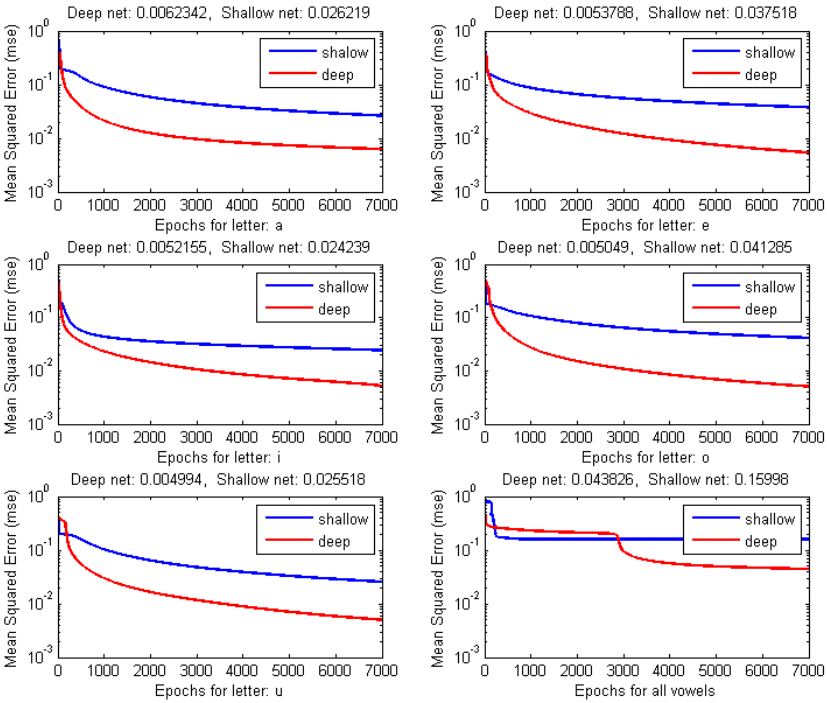


Figure 3-81. Training performance of five binary deep networks and one multiclass deep network. In this figure shallow (200,100, 2 neurons) peers of these six networks are also shown.

In Figure **3-81** we can make multiple observations. Firstly, in all cases (binary and multiclass) deep networks trained with deep learning show better performance than shallow networks trained with standard backpropagation. Secondly, in average five binary networks perform better than a single multiclass network in both cases deep and shallow. Further, five binary networks either shallow or deep, perform better in average than the two multiclass networks (deep and shallow). Therefore, the best classification of vowels in training is achieved by five binary deep networks. Similarly, the second better classification is achieved the five binary shallow networks. In Figure **3-82** we show the precision-recall validation curve for these 10 best networks.

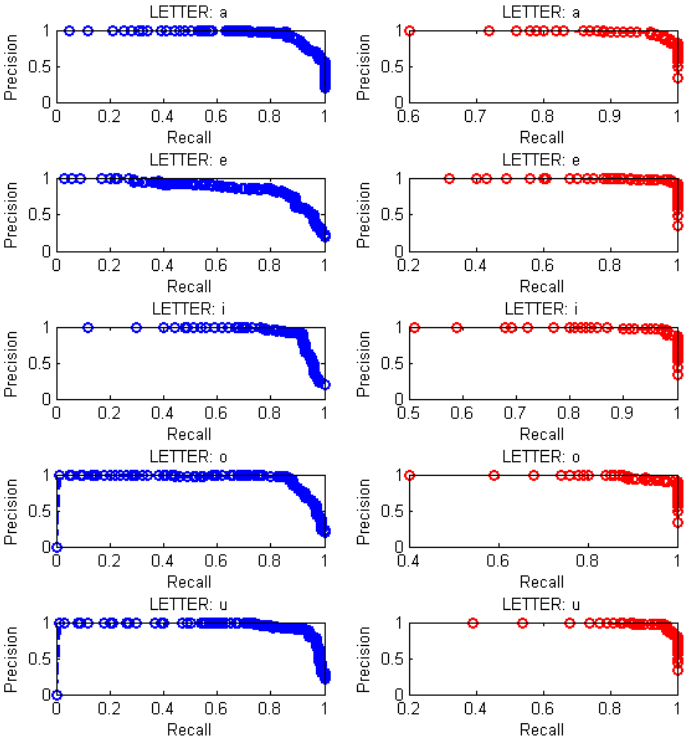
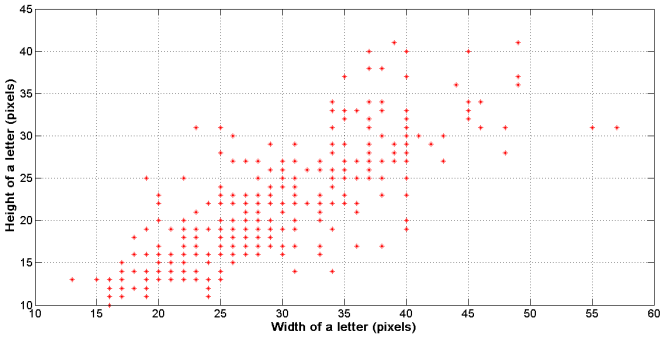


Figure 3-82. Precision-recall curves for validation of five binary deep networks and five binary shallow networks. As expected and reflecting the results in training performance, deep networks are fairly better than shallow ones.

In Figure **3-83**, we can view a random example of our trained networks processing one image. For the first network we do not slide the 20x20 window through each pixel of the picture, as it slows down unacceptably the detection. Rather, we slide the window each 20 pixels; this improves the searching to reasonable time, yet not real. The pixels detected as positive by the detector (i.e. there is text), are then dilated to form the actual textual areas (Figure **3-83**). Across these textual areas, we then run our second network as a text recognizer. Oppositely to the previous one, we have to slide several windows, for, as before mentioned, we don't know in advance the possible size of vowel to be detected, if any. In Figure **3-84**, we

have plotted the height and the width of many of the vowels collected manually during the training phase, before they were normalized to 16x13 pixels. Notice that letters not taken into training are expected to exhibit nearly these sizes. In theory, therefore, we would need a sliding window for each size in order to detect them all. In practice, however, we can find the most popular or representative sizes. We find the most representative sizes by applying a k-means clustering algorithm (of 4 classes) in the space of windows sizes (Figure **3-84**). The centroids of these classes are chosen as the sizes for our sliding windows: 15x20, 28x21, 36x47, and 44x24 pixels. Finally, 4 is a number assumed just to keep the algorithm runnable in standard computers; it can be higher though.

Figure 3-83. A frame processed with our detector (top-left). The detected points dilated in a binary image that shows the textual areas (top-right). The two previous images merged in one (bottom-left). The result of our "a" classifier applied across the white textual areas (bottom-right). Note that in its majority, false positives of our detector are finally filtered by our second network.



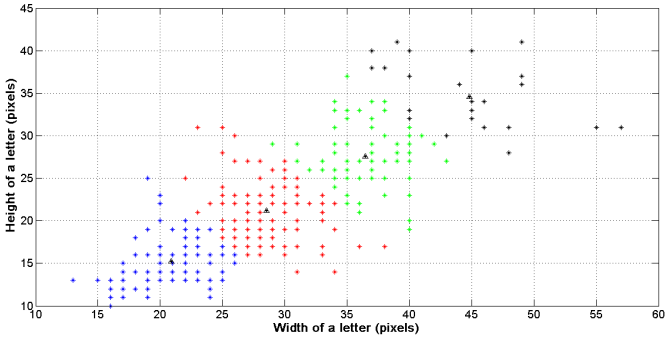


Figure 3-84. On top of this figure the sizes (height and width) of the letters selected manually during the training phase are shown (only 400 examples are plotted). In the bottom figure, a 4 class k-means segmentation algorithm has been applied to these points. Classes are labeled with colors and centroids with a triangle. These centroids are assumed to be the most popular sizes for expected letters. Thus, we fix 4 sliding windows for our second network to be these sizes.

Does it work for See ColOr?: Our approach is not real time as doesn't it the state of the art. Importantly though, real time for See ColOr was not intended in text recognition. As presented at the opening of this section, for blind individuals there is a way to maintain a real time text-reading application, without going to the trouble to scan a whole image. See ColOr already has a tactile interface that serves our global module, so it will serve too as a platform for text reading. Rather than analyzing the entire image looking for text, we display the image into the tablet and have the user scan it himself with his fingers. The overall strategy is: *letter touched, letter told*. Users then will receive textual feedback using their fingers, quite like braille system. Except that in See ColOr the feedback is acquired through the auditory pathway rather than the tact. In practical terms, it is like the user has a sliding window stuck in his fingertip as it moves through the tablet, so we can focus our networks on that window only. More precisely, five windows are to be evaluated: one for the detector and four for the recognizer (Figure **3-84**). Moreover, as it has been studied before in this thesis, the use of the protocol TUIO (Tactile interfacing) allows us to communicate the coordinates of the finger wirelessly to a computer (running the networks) from any Mac device; e.g. iPad, iPhone. We have implemented this application in an iPhone to show the functionality, portability and practicality of our approach (Figure **3-85**). A video of our neural networks working on our iPhone-based application for text recognition in See ColOr, can be watched in: http://youtu.be/zGq7KrcQ0Ks

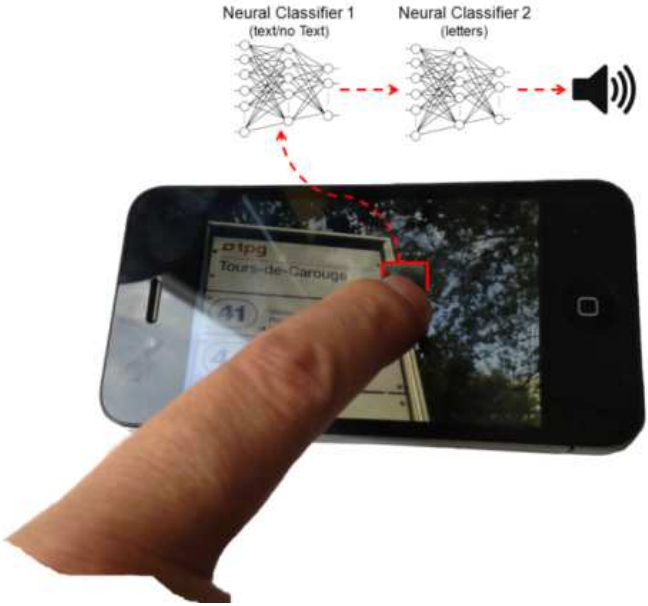


Figure 3-85. Our implementation of braille-like text recognition, using deep neural networks, deep learning and iPhone. First a neural text detector decides whether or not the finger tap is text. If so, a neural text recognizer decides what is the letter touched by the finger.

Testing accuracy with real scenario data: As it was already mentioned, we trained our system with three videos recorded in three different bus stops of Geneva, namely: Tours-de-Carouge, Grange-Collomb (both in the area of Carouge) and Jonction in the city area. For testing, we recorded two new videos in yet another two stops, namely: Gare-Cornavi in downtown and Petit-Lancy in the periphery. The geographic locations of these places within the city of Geneva are shown in Figure **3-86**. Out of these two testing videos we manually extracted 100 examples per vowel with their respective coordinates within the image. Afterwards, 500 examples of background were extracted randomly (but always outside of textual areas) and saving their coordinates within the image.

Figure 3-86. The bus stops used for this test in Geneva.

Since all collected coordinates belong to any of six classes, i.e. they belong either to one of the five vowels (textual), or to the background. To test the accuracy of our approach, we just need to evaluate the corresponding networks in these coordinates. In the one hand, to test our first network we extract a 20x20 patch of a given background-class coordinates. In the other hand, our second neural network is fed and tested with four patches (15x20, 28x21, 36x47, and 44x24 pixels) at each of the textual coordinates. The answer of the network is positive if at least one of these four inputs is positive, i.e. a letter was found at some scale or size. The results yielded by this experiment are reported in Tables Table **3-6** to Table **3-9**.

	True Pos.	False Pos.	False Neg.	Precision	Recall
A_net	94	3	6	0.969	0.940
E_net	97	2	3	0.980	0.970
I_net	98	9	2	0.916	0.980
O_net	100	7	0	0.935	1.000
U_net	96	6	4	0.941	0.960
Text_net	483	73	17	0.869	0.966

Table 3-6. Test performance for multiple binary networks trained with deep learning.

	True Pos.	False Pos.	False Neg.	Precision	Recall
Class A	92	4	8	0.958	0.920
Class E	90	5	10	0.947	0.900
Class I	93	7	7	0.930	0.930
Class O	96	9	4	0.914	0.960
Class U	94	9	6	0.913	0.940
			Multiclass Net=	0.932	0.930

Table 3-7. Test performance for a multiclass network trained with deep learning.

	True Pos.	False Pos.	False Neg.	Precision	Recall
A_net	90	13	10	0.874	0.900
E_net	92	7	8	0.929	0.920
I_net	86	16	14	0.843	0.860
O_net	83	11	17	0.883	0.830
U_net	94	10	6	0.904	0.940
Text_net	412	73	88	0.849	0.824

Table 3-8. Test performance for multiple binary networks trained with standard backpropagation.

	True Pos.	False Pos.	False Neg.	Precision	Recall
Class A	87	12	13	0.879	0.870
Class E	87	14	13	0.861	0.870
Class I	82	18	18	0.820	0.820
Class O	79	16	21	0.832	0.790
Class U	81	20	19	0.802	0.810
			Multiclass Net=	0.839	0.832

Table 3-9. Test performance for a multiclass network trained with standard backpropagation.

Reflecting their best performance in training, five binary deep neural networks were the best architectures in this test. Therefore, we could conclude that having a deep network per letter of the alphabet, is the best strategy for a full text-reading implementation. That would be true were it not for each letter we add increases the time response of the system, which is proportional to the number of networks analyzing the finger position. Oppositely, if we chose to have a multiclass network that classifies all the letters at once, this time would remain constant since all the analysis relies on one net. This latter choice will, however, diminish the precision or accuracy of our system, as reflected in both training and test. Thus, we are left with a balance decision between precision and time. Therefore, the multiclass network option

must be by no means rejected. Though if we choose so, deep networks should be used rather than shallow ones.

Testing usability with real scenario data: For this test we trained our system to recognize digits from 0 to 9, exactly in the same way we did with the vowels case. The aim here is to have blindfolded people discover the number of the buses in a random station's picture. To do so, test takers need to trigger the sounds of the numbers with their fingers, while scanning the screen of the iPhone (where the image is represented but not actually shown). More precisely, every time they touch or tap a digit, our system speaks out the digit itself, allowing people to understand which and where the numbers are. To complete this task, every participant is trained with one only picture. Oppositely to the actual test, during this one-image training they are allowed looking at the picture that is shown in a computer (Figure **3-87**). For we want them to get familiar with the precision, sensibility and speed they need to apply in their fingers to make the system respond properly.

Figure 3-87. Training phase for this experiment. Computer and iPhone communicate each other. At the moment this picture was taken, our system spoke out the number 1.

Afterwards, they turn their backs to the computer and get blindfolded to initiate the test with a different image (Figure **3-88**). Importantly, the blindfold guarantees they cannot see their fingers either. Seeing the fingers move within the iPhone screen makes quite a difference, even if the screen hides the image. Finally for this test, even though images come from different bus stations, they all have only 3 bus numbers (6 digits). Thus, this is the maximum of buses that participants are to find to make the test come to an end.

Figure 3-88. Some participants taking the test. Searching the bus numbers.

We conducted experiments with a total of 17 people. Only two of them couldn't conclude the test successfully: the first one just gave up the search after 4 minutes. As for the second one, though he found all the digits separately, he couldn't assert the actual number of the buses. The performances in this test for the rest of the participants are described in Figure **3-89**. All of the participants showed enthusiasm for this test and likened it with a challenging game. As a matter of fact, we only tried one image per participant but, most of them manifested feeling challenged to take the test once more. Thus, they did have another chance, though this time they weren't blindfolded and they could see their fingers moving through the iPhone. This latter trial served us to compare performances and make conclusions.

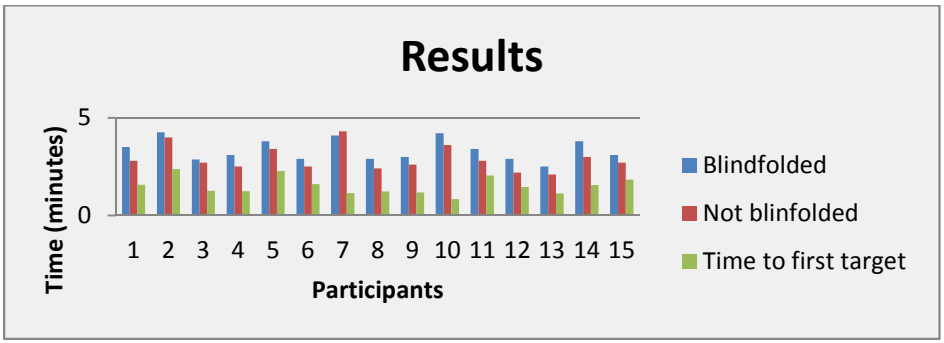


Figure 3-89. Results of this experiment.

Valuable observations can be made out of these results (Figure **3-89**). First of all, participants found the three buses in 3.35 minutes in average. This is an acceptable time for usability given that: they only had one image to train themselves for this test and get familiar with the system. This image was visible to them all the time (Figure **3-87**), which means they actually couldn't train themselves blindly, before undergoing the test. As it is for all new

systems and especially for those pursuing visual assistance, extensive training ends up in fairly better performances. Particularly, we argue this way saying that the worst five results (i.e. above 3.5 minutes) and also the two persons that failed the test were not iPhone users. In fact, those who owned an iPhone (or tactile phone) manifested being more familiar and hence prepared, to use the system than those who did not. This was very well reflected in the results. Also, when they were not blindfolded the average time lowers to 2.86; for they were more aware of the movement of their fingers by looking at them. This awareness is expected to increase as they practice to use the screen without seeing, quite like the blind do. Finally, the best results (below 3 minutes) were achieved by those who took an ordered searching strategy such as vertically and horizontally zigzag (Figure **3-90**). Those who performed the search just randomly were not as good as the former ones (Figure **3-90**). This also reflects that the correct use of tactile screens increases the success of our system. This correct use is only achievable through practice and training. Last but not least, all participants took in average 1.51 minutes to find the first number; this is almost half of the total time. In other words, once the found a point of reference the search was rather easy. This opens alternatives to improve our system such as finding at least one digit automatically to orient the user while keeping the system reactive enough.

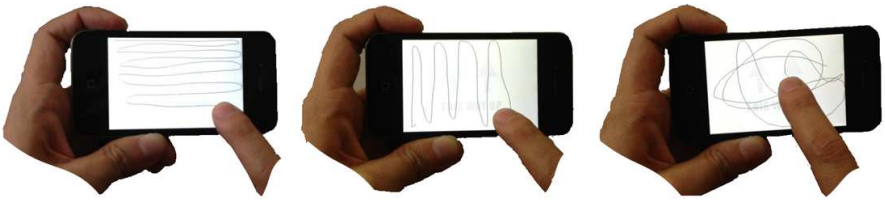


Figure 3-90. Searching strategies assumed by the participants. Left, horizontal zigzag. Middle, vertical zigzag. Right, random search. It was our observation that both zigzag-based searches were more efficient.

4 EXPERIMENTS

In this chapter we report on the most relevant experiments carried on in the course of this thesis (4.1 Past experiments). There are key experiments belonging to the earliest stages of See ColOr, which are still worth being described first in this chapter. Following, we present a set of more recent experiments conducted with both, blindfolded sighted and blind individuals (4.2 Experiments with blindfolded sighted individuals and 4.3 Experiments with blind individuals). This latter group belongs to a rural community in a developing country (Colombia, South America). Overall, experiments reported in this chapter lead us to think of See ColOr as a functional SSD, and the discussion that closes this chapter (4.3.6 But is See ColOr functional so far?) will probably prove us right.

4.1 Past experiments

In preliminary stages See ColOr underwent many tests with users; some of them are worth mentioning here. The following experiments we will comment in this section are described in detail within the article introduced by Bologna *et al.* [**229**]. The first one had as purpose to investigate whether individuals can learn associations between colors and musical instrument sounds (From colors to instruments sounds in See ColOr). Another topic of investigation was whether it is possible to interpret pictures. Several experiments were carried out by six blindfolded sighted subjects, listening to the sounds via headphones. In these experiments we used the T3 tactile tablet section (see Evolution). Participants were trained to associate colors with musical instruments and then asked to determine on five pictures (see **Figure 4-1**), objects with specific shapes and colors. Experiments involved a training phase with the use of elementary pictures. For all our experiment participants, training lasted about 45 min. [**229**]

Figure 4-1. (taken from [229]) Images used in the experiment for color-instruments association in preliminary stages of See ColOr.

Several observations were made in this experiment, namely: Regarding the children drawing illustrated in **Figure 4-1** (leftmost), all participants interpreted the major colors as the sky the sea and the sun; clouds were more difficult to infer (two individuals); instead of ducks, all the subjects found an island with yellow sand or a ship. As for the dolphin drawing,

all participants interpreted the major colors as the sky and the sea; an individual said that the dolphin was a "jumping animal", another said that it was a fish and the others determined a boat or a "round shape"; only a person found birds and the small fish was interpreted as a rock by two persons. On the interpretation of real images, such as the picture shown in **Figure 4-1** (middle), four participants correctly identified the tree with the grass and the sky; a participant qualified the tree as a strange dark object and finally, the last individual inferred a nuclear explosion. Concerning the rocky mountain (**Figure 4-1**), all subjects found major colors (blue and yellow); however, no one made the distinction between the sky and the sea. Note, however, that two participants suggested that the gray/white area between these two components represented clouds. The last assignment was to find a red door in **Figure 4-1** (rightmost). The two red doors represent less than 1% of the picture surface. **Table 4-1**, summarizes the time durations for the exploration of pictures. [**229**]

Participant	Figure 4-1 **(leftmost)**	Figure 4-1 **(left)**	Figure 4-1 **(middle)**	Figure 4-1 **(right)**	Figure 4-1 **(rightmost)**
P1	8.3	6.7	9.1		8.7
P2	7.0	8.5	7.5	7.3	4.9
P3	9.7	13.5	9.0	11.0	6.0
P4	9.2	11.3	8.7	9.0	4.8
P5	14.3	10.5	5.4	5.6	6.0
P6	9.4	11.2	10.0	10.0	9.0
Average	9.7±2.5	10.3±2.4	8.3±1.6	8.6±2.1	6.6±1.8

Table 4-1. (taken from [229]) Time results (minutes) of the experiments on color-sound association.

Another relevant experiment was "pairing colored socks". The purpose was to verify the hypothesis that with the use of a camera, it is possible to manipulate and to match colored objects with an auditory feed-back represented by sounds of musical instruments. Participants were not asked to identify colors, but just to pair similarly colored socks. The experiments were performed by seven blindfolded adults who were not present in the previous experiments. The training phase includes two main steps: associations between colors and sounds and learning how to point the camera toward the socks. Five pairs of socks having the following colors were used: black, green, low saturated yellow, blue and orange. **Table 4-2** illustrates the results of our experiment participants and **Figure 4-2** shows an individual examining a sock. It is worth noting that the average number of paired socks is high. Participant P4 made a mistake between yellow and orange socks. This experiment showed that blindfolded individuals can manipulate objects by pointing a camera on them and also that

five colors can be matched with high accuracy even after a short training session. Note that the experiment was difficult for our participants, since the camera was above the eyes [**229**]. The problem underlying was that as sighted persons they unconsciously tended to target the objects with the gaze, so their eyes served them as reference point. Shifting this reference point towards the forehead turned out to be confusing, since the target kept on going out of focus. In the final chapter (Discussion) of this thesis we will make further reference to this topic.

Figure 4-2. (taken from [229]) A participant taking the "paring colored socks" experiment in the preliminary stages of See ColOr.

Participant	Time (mn)	Success rate (pairs)
P1	16	5
P2	4	5
P3	18	5
P4	6	3
P5	15	5
P6	11	5
P7	7	5
Average	11±5.5	4.7±0.8

Table 4-2. (taken from [229]) Results of the experiment "paring colored socks".

The last experiment we want to mention is "following a colored serpentine". This experiment was quite relevant as it is one of the first attempts to autonomous mobility using See ColOr. The purpose was to verify the hypothesis that it is possible to use the See ColOr interface to follow a colored line or serpentine painted on the ground of an outdoor environment. **Figure 4-3** illustrates an individual performing this task. For this experiment we included

the same seven individuals who carried out the experiment on colored socks and additionally a blind person. The camera here was the Logitech Quickcam Notebook Pro. The training phase lasted approximately 10 min. Specifically; a supervisor managed an experiment participant in front of the colored serpentine. The experimenter was asked to listen to the typical sonification pattern, which is red in the middle area (oboe) and gray in the left and right sides (double bass) (see From colors to instruments sounds in See ColOr). Afterwards, we asked the participant performing the experiment to move the head from left to right and to become aware that the oboe sound shifts (the red line). Note that the supervisor wears a headphone and can listen to the sounds of the interface. Finally, the experimenter was asked to start to walk and to keep the oboe sound in the middle sonified region. Note that the training session was quite short. An individual had to learn to coordinate three components. The first was the oboe sound position (if any), the second was related to the awareness of the head orientation and the third was the alignment between the body and the head. Ideally, the head and the body should be aligned with the oboe sound in the middle. **Table 4-3** summarizes the results of this experiment. [**229**]

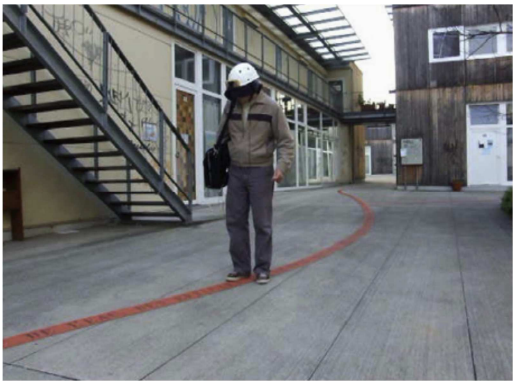


Figure 4-3. (taken from [229]) A participant taking the "following a colored serpentine" experiments in the preliminary stages of See ColOr.

Participant	Path length (m)	Speed average (m/h)
P1	88	723
P2	84	710
P3	110	485
P4	93	656
P5	84	484
P6	97	600

P7	97	451
Average	93.3±9.2	587.0±114.1

Table 4-3. (taken from [229]) Results of the experiments "following a colored serpentine".

4.1.1 Discussion

These experiments were important as they gave us an idea of the order of magnitude of the slowdown factor related to the substitution of the visual sense by the auditory pathway. In the static pictures exercise, for instance, one could say that this factor might correspond to the order of magnitude related to the ratio of visual channel capacity to auditory channel capacity, which corresponds to two orders of magnitude [**6**]. Whereas for the "socks" and the "serpentine" exercise this factor could be equal to one. From a cognitive perspective this would be consistent with the fact that these two tasks are simpler than the interpretation of image scenes.

Also, the results derived from the "serpentine" experiment were very encouraging, since this was our first attempt to use See ColOr as a mobility aid. In vision substitution, the behavioral criterion establishes that if a person can carry out normally the functions ascribed to vision, the sensory substitution indeed resembles vision [**14**]. At large, during this experiment unsighted individuals performed in moderate time a visual task hardly achievable otherwise (walking across a twisting path).

However, the experiment also revealed that for general navigation to be performed autonomously, it is needed more than mere transduction of low level visual features into sound. In particular, after this experiment, we began to imagine the presence of obstacles with an experimenter trying to estimate the distance separating him/her to an obstacle without touching it, and the nature of the objects nearby. Moreover, we also started to plan experiments, for which depth represents an important parameter.

The analysis of static pictures also served to point out the inability of participants to make sense of an image out of the sonified representation of colors (low level visual features). While very general aspects are barely attainable, cognitive aspects which often determine regions of interest were completely neglected. This stresses again the need to implement automatic processes to deal with the analysis and interpretation of complex and large amounts of visual data, which are hardly transmittable through the auditory pathway (i.e. computer vision methods, Computer-vision-based visual substitution).

Importantly though, the experiment involving the static images and to some extent the matching of socks, showed that See ColOr's sonic code is fairly learnable (From colors to instruments sounds in See ColOr). With moderate time users were capable of mastering the relations between colors and instruments sounds so as to perform these experiments.

4.2 Experiments with blindfolded sighted individuals

We conducted studies featuring 12 blindfolded participants to prove that our system does increase the spatial awareness and the intelligibility of the environment through non-visual cueing. Specifically, we address the feasibility of haptic and auditory trajectory playback as a method to substitute visual and spatial cues of a real environment in the interest of navigation. The goal behind this question is to enhance the intelligibility of the environment by representing relevant cues through non-visual methods. To better understand how our interface enhances the intelligibility of the environment, we conducted studies on specific cases in which for unsighted individuals the environment is made illegible. Particularly, we evaluated three situations in which unsighted individuals daily experiment a strong urge to be aided: (1) the perception of colored boundaries that is made unattainable for the blind, and the perception of physical boundaries. (2) Encountering of colored targets into an environment and finally, (3) gaining awareness of the presence of walls. All of them are cases notoriously challenging in which the blind may face embarrassing situations or experiment a strong urge to be aided. **Figure 4-4** illustrates these situations and this section is dedicated to prove how our interface helps to tackle each of them.

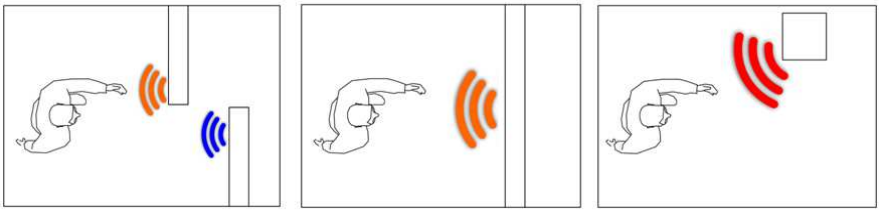


Figure 4-4. Three case studies in which our system allows the users to be aware of the close environment: (left) A user perceives a border, as with color discontinuity (timbre of sound) or depth change (rhythm of sound). (middle) A user becomes aware of the presence of a wall as he gets closer and the rhythm of the sound increases. (right) A user finds a target as the target emits a particular sound (depending on color) from a specific location in the azimuth plane.

4.2.1 Study 1: Audio Revealing of Boundaries

Figure 4-5. An individual detecting a real world boundary (in this case caused by depth discontinuity) by means of See ColOr.

This study concerns the capacity of the users to perceive, through the audio feedback, points in an image at which its aspect changes sharply or, more formally, has discontinuities. This allows the capture of important events and changes in properties of the world. Those sudden visual changes in brightness, color, depth or textures are known as edges. Edges are extremely important in visual contents as they create boundaries that define visual shapes. Visual edges encode great amount of information in natural images. In fact, most of the natural images may be still understandable despite the absence of color and many other visual cues but edges. Therefore, attainability of edges perception could easily lead the blind persons to the acquisition of objects' shape and regions boundaries.

The interface presented in this paper allows the perception of two classes of edges, those defined by depth variations (e.g. an open door frame) as well as the ones caused by color interruption (e.g. the boundary of a colored painting hanging on a white wall). Based on the theoretical ability of identifying variation on instruments sounds (color) and frequency of sound (depth), we hypothesize:

H1. *A fingertip scanning will be enough for a user to perceive an edge and its location on the tactile tablet (iPad).*

Procedure

We recruited 12 participants (7 female, 5 male, average age 28) for this study. Before this experience, participants had not used our system. After training, participants were tasked to identify five borders in each case: variation of sound timbre (color-caused edges) and rhythm of the sound (depth-caused edges). They were blindfolded and asked to answer whether or not at least one edge was present on the iPad. In the affirmative cases, they had to prove his answer by sliding their finger along the edge. In order to exanimate positive and false cases, they were placed (sometimes) in front of flat blank walls (edgeless). The answer, therefore, must be negative for these exceptional cases. For the rest and the majority of the cases, they were placed in front of a prominent border (corners, opened doors, discontinuity of color in a wall). In the interest of easing the evaluation, we granted that only one edge was captured by the camera and rendered on the iPad.

4.2.2 Study 2: Spatial Awareness

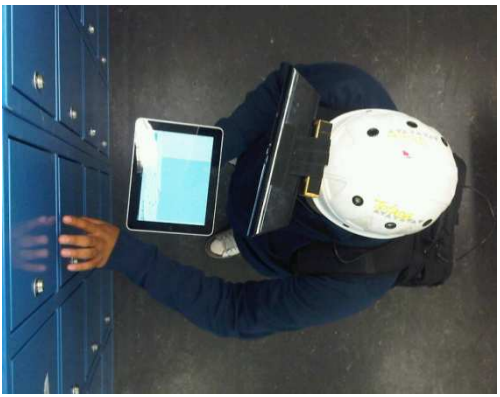


Figure 4-6. An individual becoming aware of a wall-like obstacle on his way (in this case a rack of lockers). He was able to stop himself to a distance from which the wall is reachable by hand.

This study concerns the ability to be aware of oneself in space. Specifically, in this study we investigate awareness of spatial relationships as the skill to perceive an object in relation to oneself. Spatial awareness is usually defined to include a person into space, so that a user will understand his location and the location of objects in relation to his body [**188**]. In understanding these relationships, persons come to mechanize concepts such as distance and location. For example, a person with spatial awareness will understand 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 however, entirely lack this ability and their positioning in relation with the world is a thorough trial-and-error process.

Finding an unexpected wall or encountering an unexpected closed door often becomes an embarrassing event that blind people must face daily. Based on the theoretical ability of identifying the increasing/decreasing variation of a sound's rhythm (depth), we hypothesize:

H2. *Using our interface, a blindfolded person walking toward an obstacle will be able to stop just before hitting it. In fact, this can be done so accurately that (s)he might know whether the object is already reachable by hand.*

Procedure

This study featured the same participants as the previous one. They were blindfolded again and given training. As a consequence, they were able to accurately identify the rhythm of the sound that a wall produces in our system according to its nearness. Afterwards, they were five times asked to walk down the corridor (from different distances) toward a wall and stop right before a hit, but close enough to reach it by hand. Once stopped, the participants were asked to reach out and touch the wall.

4.2.3 Study 3: Finding targets and detecting obstacles

Figure 4-7. An individual with a tactile tablet seeking a person that wears a red t-shirt by means of See ColOr.

The final study aims at evaluating capabilities of individuals to seek and find a specific target disposed into the environment being explored. Locating something in our surrounding is somehow a constantly happening labor that we perform with multiple purposes such as, reaching a desired venue, finding an object, avoiding collisions, making ourselves aware of the surroundings layout and in general terms, it greatly supports exploration as well as navigation tasks.

When sighted individuals are asked to indicate "where is something?", they first perform a gaze scanning of the surrounding and once localized, they point out the target. The pointing operation in this case refers to the designation of that object by mediation of the arm, finger and sight; aligned sight and finger form a "natural pointer" [**129**]. Our interface preserves this mechanism as much as possible: We substitute the gaze scanning by a global exploration with their fingers. The natural pointer, in turn, continues to be one of the fingers that stops on the area the object is thought to be. Our hypothesis in this study states the following:

H3. *Through the use of our interface the blindfold user will recover the ability of seeking and finding an entity located into the environment.*

Procedure

This study featured the same participants as the previous ones. A person acting as a target and wearing a red t-shirt stands in front of the blindfolded user at an unknown location. The blindfolded person is then tasked to explore with his fingers the whole panorama being captured by the camera and rendered onto the iPad. We fixed a controlled environment so that no red elements others than the t-shirt are present. This simplifies the task into detecting which portion of the tactile tablet screen emits the sound of red when touched. This trial keeps on being run five times as the target person moves arbitrarily.

4.2.4 Results

This section reports numerical results achieved in this work concerning the case studies previously described. It is worth noticing that the study on edges was twofold: Perceiving colored edges (changing sounds) and perceiving edges of depth (variation on rhythm of sound). We thus report results individually for the two of them. The apparatus used in these experiments were: one touch-pad (iPad or tablet), one helmet-mounted camera range (Microsoft Kinect), one 14" laptop carried on a haversack and one set of high quality headphones. **Figure 4-8** plots four matrixes that graphically summarize the trend of each of the studies carried out, respectively. Rows of these matrixes represent the number of trials or repetitions (5) in which each study was consecutively performed. Columns, in turn, represent the participants who performed the study (12). To better understand the meaning of these matrixes we describe them separately:

For the matrix perceiving colored edges (Top-Left **Figure 4-8**), the green color means that the participant perceived the edge when he first touched it. Yellow, in turn, means that the participant required touching the edge more than once in order to perceive it as such. Finally, the red color means that the participant either failed to perceive the border or perceived it wrongly situated (deceptive). The same colors meaning can be extended to the matrix perceiving edges of depth (Top-Right **Figure 4-8**).

For the matrix walking toward a wall (Bottom-Left **Figure 4-8**), the green color means that the participant timely stopped few steps ahead the wall, and managed to touch it upon request. Yellow, in turn, means that the participant did stop and did not hit against the wall but, he failed to touch the wall as he stopped to walk prematurely. Finally, the red color means that it was required our intervention for the participant not to hit against the wall otherwise, he would have collided.

Finally, for the matrix seeking a target (Bottom-Right **Figure 4-8**), the green color means that the participant detected the target when he first touched it. Yellow, in turn, means that the participant required touching the target more than once in order detect it. Finally, the red color means that the participant either failed to detect the target or detected it wrongly situated (deceptive).

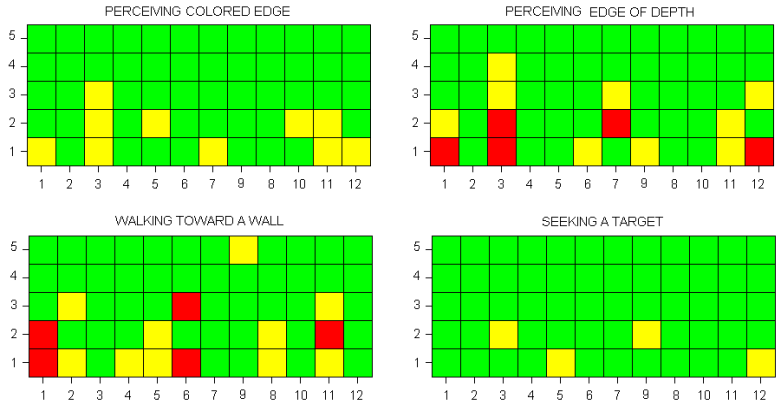


Figure 4-8. Four matrices representing graphical results of our case studies: green means goal fully achieved, yellow means goal partially achieved and red means goal unachieved. Vertical axes represent the number of trials which every case was performed in and horizontal axes represent the number of participants who took part of the study.

Moreover, **Figure 4-9** reports statistical results of this study in a global context. Each experiment represented across the horizontal axis, was performed five times by 12 participants. We thus have a total of 5x12=60 trials (100%) represented across the vertical axis. **Figure**

4-9 does not discriminate against participants' individual performance, yet it relates the percentage ratio of failures and successes, between the four distinct case studies. The colors of the bars have the same meaning as in **Figure 4-8**.

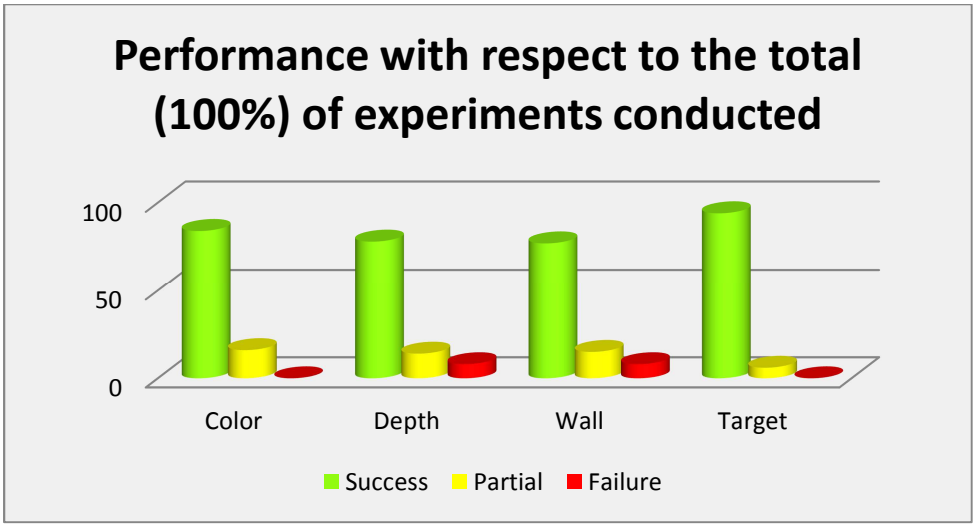


Figure 4-9. Percentage ratio of failures and successes, between the four distinct case studies.

4.2.5 Discussion

Overall, the results of our user studies (with an average of success of around 82%, **Figure 4-9**) reveal that our system aimed at fostering spatial awareness and increasing the legibility of the environment can help unsighted persons to achieve a mental map of relevant aspects of unfamiliar locations, a first step toward a robust navigational experience. All participants found the system useful to roughly grasp a layout sketch of the scene and discover entities hovering nearby. In particular, the spatial awareness evaluated in Study 2, added to the alerting system, enables unsighted individuals to travel safely across the environments as they can plan the path to a target without stumbling. Study 3, in turn, grants them access to information about resources in the environment that they may not otherwise have received.

Through observation of the participants performances it became clear the impact and importance of prior training to succeed the experiments. As proof, we can rapidly realize that the four matrixes in **Figure 4-8** pile up all non-green data (failures and partial successes) at their bottoms (first trials). Tops of the matrixes however, remain largely green as users had already gained experience by the time they performed ultimate experiments. Although this

study did not explicitly encourage entire learning of our sonic code; this is the principal component of our approach and only by full achievement thereof our system becomes truly handy. The results reported in **Figure 4-8** however, confirm that the rate of training demanded is rather low and in some cases negligible. The five minutes training prior the experiments plus the first trials of the test, seem to be enough for a user to get acquainted with system and succeed. In particular, the outcomes achieved in Study 2 (Bottom-Left **Figure 4-8**) reveal the major percentage of failures and partial successes. Curiously, most of these flaws that relate dangerous situations (hit against the wall) were caused by persons who confessed to lack nerve rather than training.

Our findings in Study 1 (specifically colored edges) and Study 3 strongly verify the hypothesis *H1* and *H3*, since no failure was observed and provided that partial successes are still good and seldom occurred in either case. Our system also reveals features that in turn, increase the intelligibility of the environment for the blind by disclosing significant information they could not otherwise perceive i.e. colored edges and target situation. Barrow *et al.* [**230**] emphasizes the importance of edges in making an image intelligible far more than other features. In fact, we argue that edges are irremovable when images are required to preserve sense. A simple exercise of blurring an image using a filter can show us that the more blurred the image the less sense it makes as its intelligibility little by little fades away. It is well known that blind people entirely lack abilities to identify color-caused edges that define objects, entities and natural boundaries. Therefore, we consider as crucial to move toward letting unsighted individuals gain this ability through non-visual cueing.

Moreover, grasping a target when there is no telling of its location happens to be a thorough situation in which unsighted people feel a strong urge to be aided. Hence, the localization of objects using our system offers a greater sense of independence and could be of great help to understand the layout of the environment. The results reported in **Figure 4-8** (Bottom-Right) clearly indicate the benefits which our system contributes to this matter with. In line with Study 3, our alerting system also adds a strong component in this aspect, since detecting obstacles/objects in natural scenes is a critical functionality in many mobility aid applications. However, a higher level module aimed at recognizing the objects after detection, would be of great help when navigating, exploring and understanding the environment. For instance, alerting someone could be done far more effectively if in lieu of launching a warning about a potential obstacle, the system notifies the nature of the object e.g. this is a man or, a garbage can. This would bring us very close to a robust visual substitution system and that is actually the venue our project targets to.

In general terms, with the addition of our haptic-based interface, as well as the alerting system we expect trained individuals to move independently in unknown environments [**30**]. Moreover, we are aware that the See ColOr interface will not allow trained individuals to understand or to discern human expressions. However, specialized modules could be developed for these challenging tasks. Finally as discussed by Bologna *et a.l* [**30**], cutting edge technologies like neural implants or retinal electrodes are increasingly improving and in a few years they might be used as local perception vision aids. It will take too long however,

before these sorts of sophisticated technologies become affordable for average people especially in developing countries. Besides, the clinical risks and physical adjustments to which users may be compromised to adapt these aids. Moreover, one arising question regarding retinal electrodes is whether an individual with implanted electrodes could be able to determine depth and color, the first parameter being absolutely crucial for mobility. We hence, believe that more financial efforts should be engaged in non-invasive mobility interfaces for vision substitution. The blind community will directly benefit from this research topic in the short/mid-term.

4.3 Experiments with blind individuals

Another key aspect of See ColOr is that the current prototype is made up of relatively affordable technologies. This will benefit low-income countries, where blindness indexes continue to grow due to poor medical accessibility in rural areas. This also makes See ColOr more practical and better situated in terms of social impact. Thus, we traveled to Colombia South America, as we believe that it is always incumbent on a researcher to make the effort to reach out to the community to engage with the population they're serving. With the support of "*Pacto para la productividad*", a Colombian government initiative[17] towards work inclusion for individuals with disabilities; we conducted over 180 experiments with 15 blind individuals, during more than 60 hours (see **Figure 4-10**). These subjects (many congenitally blind) were legally blind individuals, meaning they have visual acuity of less than 20/400. Their ages ranged between 25 and 50, and all of them had educational level above high school.

It still remains to be seen whether the sonifications of visual cues could be interpreted with sufficient accuracy, enabling the perception of environmental features in moderate time. Thus, we address here the usability of See ColOr to substitute visual and spatial cues of a real environment in the interest of navigation. Overall, we want to evaluate the proficiency of our system in guiding blind individuals with effectiveness. To obtain quantifiable and repeatable results, we conducted studies on four specific experiments representing pragmatic situations. By and large, for experiments reported in this section the results reveal that See ColOr is learnable, functional and provides easy interaction. In moderate time, participants were enabled to grasp visual information of the world 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.

[17] www.pactodeproductividad.com

Figure 4-10. A collage that summarizes the See ColOr experiments conducted with blind subjects in developing countries (Colombia, South America).

4.3.1 Study 1: reaching a colored target via spatialized sound

This study concerns the capacity of blind users to perceive, through the audio feedback, salient points (colored target) in a video stream. 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.

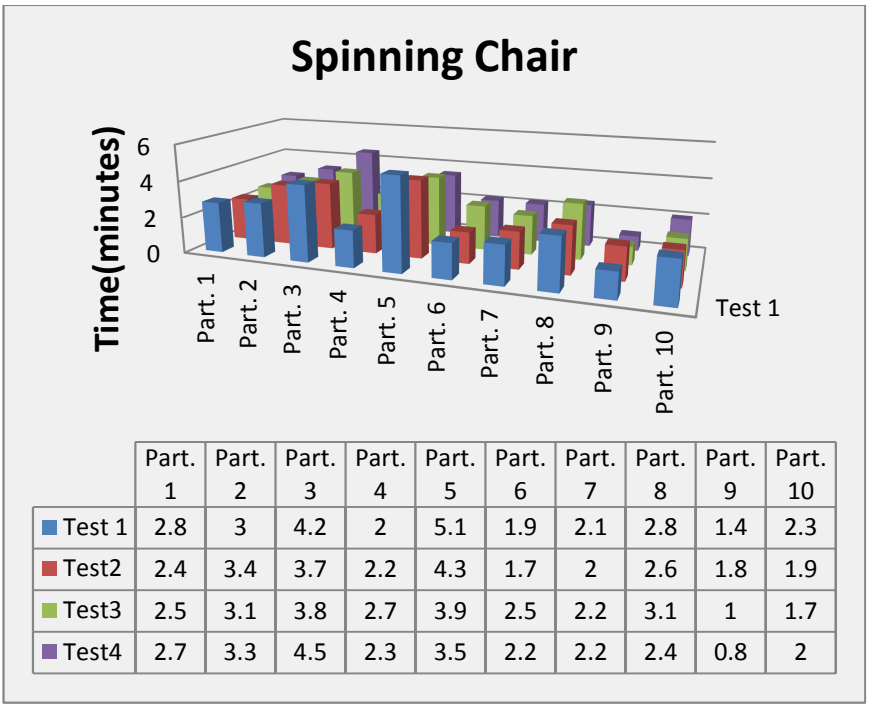
Procedure

A red target is placed at unknown location (not revealed to the participant) within a room. At the center of this room (about 2 meters away from the target), a participant sitting on a spinning chair is then asked to rotate until perceiving the target (by hearing the particular sound emitted by the red color). As the user revolves, the sound emitted by the red target may appear as coming from either right or left. At that moment, the user is asked to go on rotating slower until the emitted sound is centered in the auditory field (it is no longer heard to the right nor the left, but in the middle). Hence, the participant will know that he has rotated the chair enough to positioning himself right in front of the target. Following, the participant must stand up and walk towards the target until reach it. In this experiment the alerting system is activated, meaning that while spatialized sound leads the user to the target, the alerting system prevents him (her) from bumping into it (see **Figure 4-11**).

Figure 4-11. Participants taking the spinning-chair test.

Results

Figure 4-12 plots a panoramic bars graphic depicting the distribution of elapsed times when 10 participants completed for 4 times this experiment. Statistical analysis of these data reports that in average a participant takes 2.45 minutes to entirely fulfill the task. First quarter of these data, on the other hand, fell below 2 minutes. Also, 75% of the data never surpassed 3.2 minutes. Overall, the experiments also reported an upper adjacent data of 4.5 minutes as well as a lower adjacent data of 0.8 minutes. Finally, these data revealed an outlier value of 5.1 minutes.



	Part. 1	Part. 2	Part. 3	Part. 4	Part. 5	Part. 6	Part. 7	Part. 8	Part. 9	Part. 10
Test 1	2.8	3	4.2	2	5.1	1.9	2.1	2.8	1.4	2.3
Test2	2.4	3.4	3.7	2.2	4.3	1.7	2	2.6	1.8	1.9
Test3	2.5	3.1	3.8	2.7	3.9	2.5	2.2	3.1	1	1.7
Test4	2.7	3.3	4.5	2.3	3.5	2.2	2.2	2.4	0.8	2

Figure 4-12. Results of the spinning-chair experiment.

4.3.2 Study 2: Gaining awareness of the presence 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.

Procedure

A participant is tasked to walk (from unknown distance) towards a wall located ahead. The goal here is to perceive the distance between he and the wall as accurately as to stop at a safe distance not to collide, yet close enough to reach out and touch it. There are two strategies in See ColOr that may lead to achieve this goal:

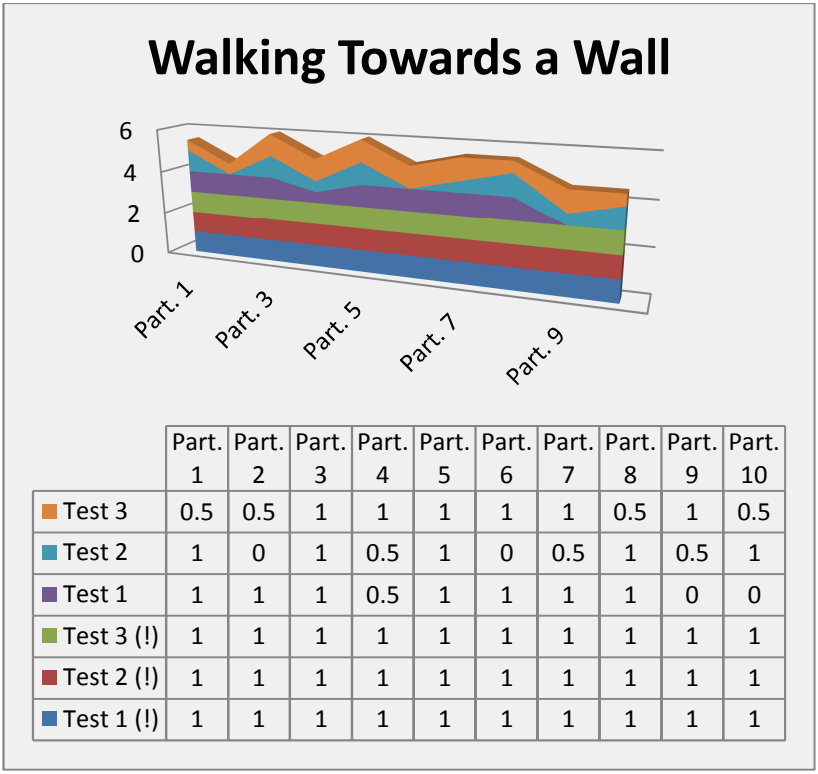
- Assessing the sound emitted by the wall itself: The closer the wall, the faster the rhythm of the sound.
- Relying upon the alerting system: The participant progresses without heeding the rhythm but, waiting for a warning being launched.

We conducted this experiment employing both tactics, just to illustrate clear advantages of using See ColOr in assisted navigation while having a functional alerting system (see **Figure 4-13**).

Figure 4-13. Participants taking the walking-towards-a-wall test.

Results

The chart showed in **Figure 4-14** reports the results attained in this experiment. In this chart the performance of participants were scored as follow: 1 means goal fully achieved (i.e. participant could stop few steps ahead, and managed to touch the wall); 0.5 means goal partially achieved (i.e. participant could stop but, failed to touch); and 0 in turn, means goal unachieved (i.e. participant required our aid not to bump into the wall). Notice that tests marked as (!) indicate they were carried out with the alerting system turned on. Likewise, tests not marked were performed as having the alerting system deactivated.



	Part. 1	Part. 2	Part. 3	Part. 4	Part. 5	Part. 6	Part. 7	Part. 8	Part. 9	Part. 10
Test 3	0.5	0.5	1	1	1	1	1	0.5	1	0.5
Test 2	1	0	1	0.5	1	0	0.5	1	0.5	1
Test 1	1	1	1	0.5	1	1	1	1	0	0
Test 3 (!)	1	1	1	1	1	1	1	1	1	1
Test 2 (!)	1	1	1	1	1	1	1	1	1	1
Test 1 (!)	1	1	1	1	1	1	1	1	1	1

Figure 4-14. Results of the walking-towards-a-wall experiment.

4.3.3 Study 3: finding, approaching and shaking the hand

Based on surveys with visually impaired and blind users, the authors in [**8**] 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; pretty much like the visual system does.

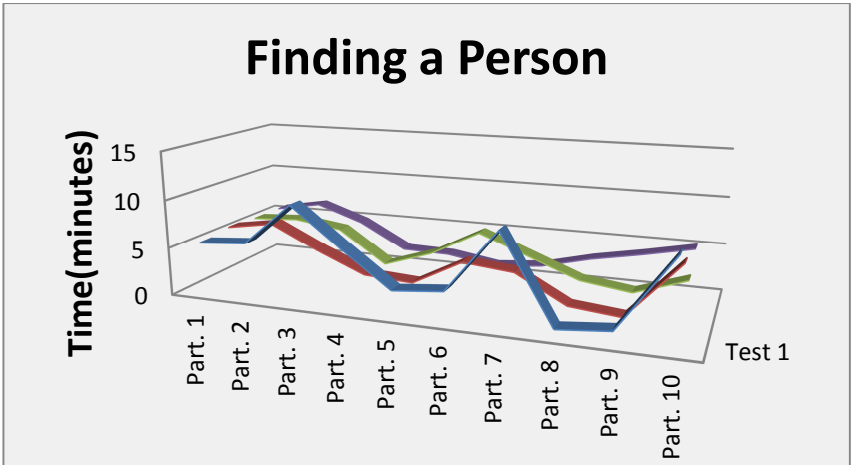
Procedure

We first learn our recognition module to identify a particular person (target). This person is requested to stand steadily at unknown location within a 15 squared meters room. Following, the blind participant is also given an initial location and encouraged to start up the search. Both positions are randomly selected in order to simulate a real situation (serendipitous encounters). The participant keeps on seeking the target as (s)he is oriented by the activation/deactivation (i.e. detection/no-detection) of the audio cue. Thus, the blind participant should be able to progressively approach the target. Once the target is thought near enough, the participant will try to shake hands. Collisions against the target or the walls, are to be eluded provided that the alerting system is turned on (see **Figure 4-15**).

Figure 4-15. Participants taking the finding-a-person test.

Results

Figure 4-16 is a mixed representation of curves describing the variation in time of 10 participants as they perform four repetitions of this experiment. Statistical analysis of these data reports that in average a participant takes 4.1 minutes to entirely fulfill the task. First quarter of these data, on the other hand, fell below 2.05 minutes. Also, 75% of the data never surpassed 5.8 minutes. Overall, the experiments also reported an upper adjacent data of 10.2 minutes as well as a lower adjacent data of 0.7 minutes. Finally, these data revealed no outlier values.



	Part. 1	Part. 2	Part. 3	Part. 4	Part. 5	Part. 6	Part. 7	Part. 8	Part. 9	Part. 10
Test 1	5.4	5.8	10.2	6.3	2.6	3.1	9.5	0.7	1.2	8.7
Test 2	5.8	6.6	4.2	2.1	1.8	4.5	4	1.2	0.7	6.5
Test 3	5.6	5.9	5.2	1.8	3.5	6.2	4.2	2.2	1.5	3.3
Test 4	5.6	6.5	4.8	2.1	2	1.2	1.7	3	4	5.1

Figure 4-16. Results of the finding-a-person experiment.

4.3.4 Study 4: grasping particular objects from a collection of items

This study concerns the retrieving of daily objects. Here we look forward to evaluating the capacity of interaction of See ColOr in orienting a user as he attempts to seize small items. For instance, when sighted individuals drop something, the regaining thereof is quite an easy task; by contrast for blind people is difficult to get into do it. In fact, fallen small objects very often yield embarrassing situations that might lower their feeling of dignity. Further, it is extremely useful to allow the visually impaired gaining awareness of daily objects they otherwise could fail to notice, or simply need others aid.

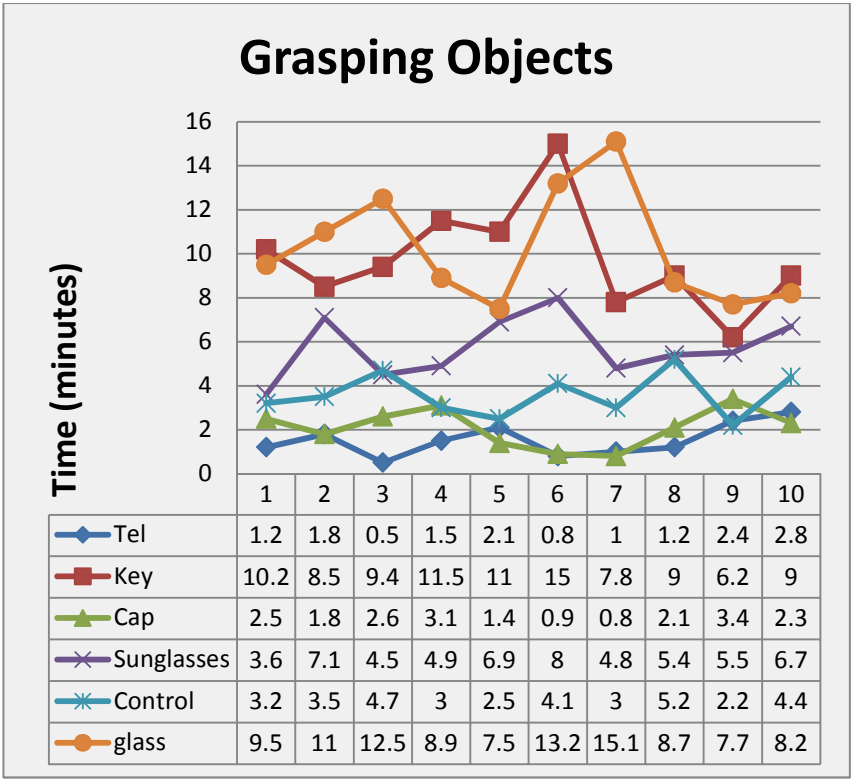
Procedure

There are six different items lying on a table, namely: sunglasses, keys, glass, cap, remote control, and a landline telephone. Also, the See ColOrs' recognition module has been trained to identify those very elements, upon request. We put blind participants to the test by asking them to find the items one by one, with the aid of See ColOr. Thus, a participant has to scan the table back and forth, while heeding the audio cues emitted by the items. When a desired item is detected, we task the user to pick it up (see **Figure 4-17**).

Figure 4-17. Participants taking the grasping-objects test.

Results

Figure 4-18 plots six curves that describe the fluctuation of time in function of 10 participants, as they grasp the six items used for this experiment in this order: telephone, keys, cap, sunglasses, remote control and glass. Statistical analysis of these data reports that in average a participant takes 1.35, 9.2, 2.2, 5.45, 3.35, 9.2 minutes to find and grasp each of the respective items. Overall, the experiments also reported upper adjacent data of 2.8, 11.5, 3.4, 8, 5.2, 15.1 minutes respectively, as well as lower adjacent data of 0.5, 6.2, 0.8, 3.6, 2.2, 7.5 minutes. Finally, the data revealed one outlier value of 15 minutes belonging to the keys search.



	1	2	3	4	5	6	7	8	9	10
Tel	1.2	1.8	0.5	1.5	2.1	0.8	1	1.2	2.4	2.8
Key	10.2	8.5	9.4	11.5	11	15	7.8	9	6.2	9
Cap	2.5	1.8	2.6	3.1	1.4	0.9	0.8	2.1	3.4	2.3
Sunglasses	3.6	7.1	4.5	4.9	6.9	8	4.8	5.4	5.5	6.7
Control	3.2	3.5	4.7	3	2.5	4.1	3	5.2	2.2	4.4
glass	9.5	11	12.5	8.9	7.5	13.2	15.1	8.7	7.7	8.2

Figure 4-18. Results of the grasping-objects experiment.

4.3.5 Discussion

Designing electronic SSDs continues to be a difficult task for a number of reasons. 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 (i.e. representation of more 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. Notice that for the visually impaired; remaining perceptual capacities are further lessened by the focus needed for the mobility and orientation tasks. Therefore, while some of these tasks may be performed in traditional ways (e.g. using a cane), further sense of independence will be attained to the extent that they are automated.

The ultimate purpose of this work was to bear out the usability of our SSD in real scenarios. We have shown that visual impaired individuals are fairly capable of interacting with our system. In other words, we introduced a functional prototype for the mobility of blind individuals. Although, there are many issues that need resolving before navigational systems can seamlessly orient the blind, our results do reveal that See ColOr could make one step further towards independent mobility thereof. It is worth also noticing that to complete these particular experiments, a user never required more than one hour of training. While mastering See ColOr will take further hours of training, this will be negligible if we consider that blind persons may spend years on learning a braille system.

Overall, experiments conducted reveal that blind users were enabled to grasp information out of which the visual world is fashioned, in moderate time. In particular, study 4 presented significant differences between finding the telephone (1.35 minutes in average) and finding the keys (9.2 minutes in average). This has to do with the recognition rate of the computer vision method that was implemented. The problem underlying here is that small items such as the keys are devoid of distinguishable features. This makes it hard for detection algorithms classifying with accuracy. Further work on feature detection and description needs to be done before reaching more consistent results. By contrast, study 2 yielded unsurpassable results after activation of an alerting system in See ColOr. 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, as least for obstacle detection. Study 1, in turn, showed that it is possible to successfully and efficiently aid a blind person in spatial orientation and obstacle avoidance through the use of spatial audio and sound guidance. Finally, 4.1 minutes in average to recognize a person (study 3) is quite an acceptable time for a blind user not to miss an opportunity to socialize and be aware of the people nearby.

Importantly, in the experiments related to the recognition module (i.e. objects and persons identification) natural speech was used to label detected targets. Though, the detection was severely 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 must be right in front). That was finally an issue reflected in the results with long searches of up to 9 minutes. The peripheral view of the camera was utterly lost, so that hundreds of detections in the lateral parts of the picture were neglected. To make it worse, in proportion to this small central area, head movements are often too fast, so that the objects appear blurred within the area. Therefore, this turns out to be a tough trial-and-error process to precisely centering an object whose position was unknown. Accordingly, in regard to the speech, 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 (see Improving time in experiments).

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. Partici-

pants stressed the fact that besides being functional enough, an aid system must be wearable and comfortable. Particularly, we highlight the request of a non-negligible number of participants about relocating the camera. Over the years, blind individuals lose the instinctive notion of targeting his head 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. Likewise, they encouraged us to make further effort to spatialize sound also in height. Participants broadly showed enthusiasm and were acquiescent for the use of See ColOr. Those who had a guide-dog, however, expressed little interest in swapping it now, unless technology makes strides rapidly enough. This clearly leaves the door open for further research and stronger efforts in pursuit of more suitable prototypes.

4.3.6 But is See ColOr functional so far?

Systems are functional to the extent that they are capable of functioning. This is to say that their design has been focused on practical ends rather than decorative or theoretical ones. Thus, key features of utilitarian systems are comfort of use and due accomplishment of the functions they were designed for. In this view, we hold that functionality in SSDs is achieved to the extent they successfully substitute a sense (substituted) for another (substituting). In the literature various criteria to assessing whether a sense is being properly substituted may be found. By and large, authors use behavioral criteria [**14**], ideal scenarios [**231**] and empirical assessments to evaluate functionality in sensory substitution. In this section we will study these aspects so as to ascribe functionality to See ColOr.

a. A behavioral criterion.

In vision substitution, this criterion establishes that if a person can carry out normally the functions ascribed to vision, the sensory substitution indeed resembles vision [**14**]. For example, Paul Bach-y-Rita [**31**] designed a device that substituted the sense of sight for the sense of touch on a subject's back. He used an old dentist chair and a camera that became the eyes, so the light pulses were enrooted to blunted needles that delivered the pattern of an object onto the back. After trials using this SSD, Bach-y-Rita claimed to have satisfied the behavioral criterion [**31**]: "If a subject without functioning eyes can perceive detailed information in space, correctly localize it subjectively, and respond to it in a manner comparable to the response of a normally sighted person, I feel justified in applying the term 'vision'." Bach-y-Rita is said to be the father of neuroplasticity and creator of the first known SSD.

Do subjects using See ColOr meet a behavioral criterion?

In experiments reported in Experiments with blind individuals, blind subjects (some congenitally blind) attained, to a large degree, spatial awareness to reach a target, abil-

ity to find and approach someone, skills to locate and seize daily objects, and independence to walk safely avoiding obstacles and walls. Note that in Bach-y-Rita's experiment mobility was not yet attempt since subjects needed to be seated on a chair keeping contact with their backs. Overall, we claim that to a great extent, See ColOr also met the behavioral criterion for these experiments. Consequently, we ascribe functionality to See ColOr in regard to this particular indicator.

b. Idealistic scenario: A system simple and effective.

Moreover, Graham *et al.* [**231**] urge researchers on visual-to-auditory sensory substitution 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 instruments to produce sounds that are by no means unpleasant. In fact, many blind subjects undergoing training with See ColOr were quoted as saying: "This is like a symphony which one feels challenged to conduct to". Note that most devices substituting vision for hearing suffer from unpleasant sound coding. In the case of the vOICe [**232**], for instance, this problem is so critical that the authors are currently in the search for more optimal image-to-sound mapping through the use of interactive genetic algorithms [**233**].

As for capturing relevant visual information in sufficient detail, See ColOr not only captures color and depth (unlike the majority) but, also reveals cognitive aspects which often determine regions of interest within a picture. At the best of our knowledge, See ColOr is unique in processing not only low-level, but high-level visual features of images in automatic mode (i.e. object recognition, face identification, obstacle detection). Furthermore, current SSDs prototypes all lack interfaces to promote proactive interaction of the user with the environment. See ColOr is the only system that offers a tactile interface (embed on a tablet) to let users make the most of the information captured by the camera. This interface is intended to grant the blind user selective exploration, discovery of points of interest, comparisons, and, in general, enjoy a greater sense of independence [**234**].

See ColOr is simpler and not necessarily less effective.

In our recognition module the sonification of virtual objects is usually achieved through natural speech (rather than 3D sonification for the other modules). Some may argue that this triggers just visual imagery rather than vision via sensory substitution. In fact, Ward *et al.* [**14**] say that while a car horn evokes the image of a car, this is very different in nature from the information in a soundscape (produced by an algorithm that maps an image 'containing a car' into sound). The horn sound turns out to be general symbolic mapping mediated by the concept 'car', whereas the soundscape may convey specific information of the scene, car's type, perspective, location and so forth. State-of-the-art

SDDs use soundscapes to represent visual images with objects, the vOICe serves as example [**78**].

We argue that, besides being quite unpleasant and extremely difficult to understand, soundscapes could not convey as much information as people think. Experiments with the vOICe [**232**] show that after training, a particular user could recognize a computer within an image, as a box-like entity devoid of details. In sharp contrast, using another image, the same participant could recognize a Christmas tree with extraordinary level of detail. The reason for this is the participant did not get to know a computer while having functional vision. We adhere to Ward [**14**] who enquires whether the vOICe's users are likely to tap their prior experience of vision to augment their "visual-like experience" supposedly elicited by the soundscape.

Furthermore, a plausible conjecture is that after 50 hours of training in recognizing the soundscape representative of a Christmas tree image, one is more likely to simply develop an associative pattern between the sound and the image, rather than having the visual experience as such. Nonetheless, Amedi *et al.* [**71**] show that after 70 hours of training in soundscape identification, blind individuals started to show little activity in their primary visual cortex (i.e. cross-modal plasticity [**5**]). However, whether this activity corresponds to actual visual experiences remains largely unknown, as it is the subject of even philosophical debates [**5**] on consciousness.

See ColOr is practical.

We advocate the use of natural speech to label objects, against soundscapes, as it prevents See ColOr's users from spending 70 hours of training (an above) in recognizing an object. The end result is a system absolutely learnable, intuitive and extremely practical. Here practical refers not only to the ease of use, but to the efficiency in handling user's prior knowledge to avoid tough learning processes. In congenitally blind we exploit the concept of the object whatever they have (e.g. tactile), and in people who became blind we exploit their visual imagery. Nonetheless, we do not rule out gaining better insight into the scene by spatializing the speech to represent left/right, varying the pitch (top/down), and adjusting the volume (foreground/background). Thus, a scene presenting various objects would convey a great deal of the information expected into a soundscape, though preserving simplicity and intuitiveness (as presented in Improving time in experiments).

c. Empirical criterion: Comfort of use.

Finally, an empirical criterion to evaluate the functionality of a SSD is that in aiding the substituted sense, the substituting sense should not be missed or diminished at all. This is an issue widespread in auditory-based substitution of vision due to the use of headsets. At large, people are reluctant to block out their ears even in exchange for assis-

tance: “I already miss a sense and wearing headphones is like taking one more way”, that is a commonplace comment among participants. Indeed, natural audio cues of the environment relevant for self-orientation are likely to be missed in this way. See ColOr copes with this drawback using bone-conducted sound. This strategy turned out of broad acceptance among users, a key aspect to add to the functional features of See ColOr. Last but not least, the use of sound in See ColOr rather than touch is intrinsically advantageous. Essentially, increasing the resolution in auditory devices is doable at level of software rather than hardware.

Despite the proliferation of SSDs, many of these devices still lag behind practical aspects. 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 [**78**] [**232**] [**231**]. 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 (e.g. interpretation of complex sounds whose lengthy calculation, besides, slows down the interaction). As proof of this, we observe in the literature the lack of experiments and testing on mobility (the most practical aspect of vision [**14**]). In this sense, experiments reported in Experiments with blindfolded sighted individuals and Experiments with blind individuals show that See ColOr is going in straight line to functionality (i.e. designed to practical ends rather than theoretical). Indeed, many aspects still need to be enhanced such as reactivity in object retrieval (see Improving time in experiments). And, of course, total visual substitution is still far from being achievable, 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. Therefore, we feel justified in saying that See ColOr is to a large extent functional.

4.4 Search optimization in our experiments

For the experiments related to the recognition module (i.e. objects and persons identification) natural speech was used to label detected targets. The detection, however, was severely constrained to the central part of the image; i.e. 20% of the picture. As a consequence, the searching times increased considerably since most of the image was left unexplored. As a matter of fact, this ended up in an issue reflected in the results with long searches of up to 9 minutes. In other words, the peripheral view of the camera was lost (i.e. tunneling phenomenon), so that hundreds of detections in the lateral parts of the picture were neglected. Reason why actual detections were difficult and hence lengthy. To make it worse, in proportion to this small central area, head movements are often very fast, so that the objects appear blurred within the area. Thus, this detection turned out to be a tough trial-and-error process to precisely centering an object whose position was unknown.

This reduction of the searching area of the image to the center was used to give the user a spatial reference with respect to the target: if detected, the target must be just in front. Otherwise, it would have been impossible for a user to know the part of the image in which an object was detected. For example: an object is being detected! but where? left, right, top, center, bottom? Not knowing the position of the object with respect to the camera, hence to the head, makes it hard for the user to walk toward it and reach it. In short, detections being reported only in the central area granted the direction (straight) toward the objective. This being said, as an alternative not to constrain the image we propose in this thesis to improve the speech label of detections by adding: spatialization of the speech to represent left/right, modification of pitch to represent top/bottom (height), and adjustment of the volume (or rhythm) to represent the depth of the objects. In this way the user will always be aware of the area of the image where the detection was done. More precisely, this gives clues to the blind individual about where to go for the target and, at the same time, prevents the system from rejecting detections. Since detections will be more likely using the whole image, the searching time is bound to be reduced. **Figure 4-19** depicts this idea of translation of visual features into audio cues.

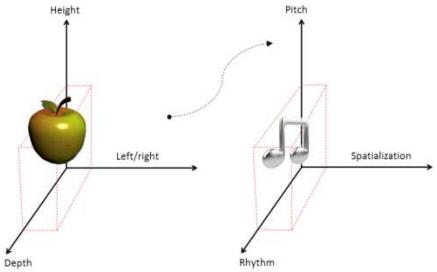


Figure 4-19. Mapping from visual hallmarks into the audio. The sound could be speech (the name of the object 'apple'), earcon, auditory icon or any other, as studied in Acoustic virtual objects.

The advantage of this strategy is that spatialization is natural for the human ears, whereas the rhythm is a very well-known technique to represent nearness, which was tested in section Study 2: Gaining awareness of the presence of walls. Otherwise, the pitch (high for top and low for bottom) is quite an intuitive method taking no effort from the user to be assimilated. In fact, we have already developed a pilot implementation of this strategy (in Matlab) yielding encouraging results. Objects were detected using our recognition module (Object recognition) and labeled with the speech 'here'. The distances of detected objects were codified into the rhythm of the speech (repetitions) as described in How does See ColOr sound like?. Then, spatialization of speech (Lateralization of sound) was used to distinguish right from left. Finally, we used an audio software app to modify the pitch of the speech in ten levels. In our laboratory, we have conducted beta tests that have yielded flawless results when participants are tasked to point out (with the finger) the position of a detected object.

Although, this approach still lacks systematic testing, some pictures displayed in **Figure 4-20** give insight into the accuracy and expediency of the method.

In **Figure 4-20**, an object is recorded by a person and then sonified. A blindfolded participant assumes his (her) own head as the center of this image, as though (s)he had got cameras for eyes. Hence, the user pinpoints the place where the object lies into the image with reference to his (her) head. The image in **Figure 4-20** shows, for example, that when the object is shot in the up-left corner (row 1, column 2), in recognition of this the user points that very corner (row 1, column 1). Likewise, images in **Figure 4-20** (row 2, column 2) describe this mechanism for an object captured at the center. Notice that in **Figure 4-20** (row 3, column 4 and row 2, column 4) even though the object stays at the center, the camera is moved away. This fact is understood by the participant who reaches out the arm, in recognition of the distance. Upon full implementation of this method in SeeColOr we expect:

- More spatial awareness for the users who will be capable of identifying and locating objects in the environment, with respect to their bodies.
- Lower time for Study 4: grasping particular objects from a collection of items, whenever the image won't suffer restrictions while being examined by our object recognition engine.

Figure 4-20. spatial-localization's beta test.

5 DISCUSSION AND CONCLUSIONS

General

We cannot stress enough the need to helping the blind and the visually impaired to gain a more independent life in a daily basis. Certainly, a great deal of potential help lies nowadays in the overlap of empirical research on sensory substitution and technology strides. This latter comprises computer science and many of its branches such as robotic, computer vision or artificial intelligence. As researchers, this calls on us to keep trying our best in breaking new ground and pushing the boundaries of the knowledge in these areas. This thesis began by offering insight into what vision/blindness implies in humans. It follows that actual vision embraces both, sensation and perception. The former more related to the sensing or acquisition of visual cues found in the outer world. Whereas the latter has more to do with the coherent interpretation of such information, allowing us to derive sense out of the world. In terms of English psychologist Nicholas Humphrey, sensation is evidently related to "what is happening to me"; while, perception is evidently related to "what is happening out there"; something way more complex. For instance, the redness as we see it arises from one's sensation, yet contemplating and understanding a rose being red is quite another thing. It is perception indeed. In the one hand, therefore, eyes and optic nerve are typically regarded as receptors that enable mere sensations. In the other hand, the brain gathers the makings of a meaningful visual experience as such (i.e. qualia), hence its association with actual perception.

The implications that follow the previous statements are such that they have given rise to the theory of sensory substitution and multisensory perception. The chief idea is that since the working of the brain is not affected in most of the cases of blindness (only the eyes), people who lose the ability to retrieve data from their eyes could still create subjective images by using data conveyed from other sensory modalities. In other words, elicitation of visual experiences in eye blindness might still be possible, provided that visual sensations somehow can reach the visual cortex of the brain. To do so, firstly, sensations from the visual space are to be mapped into another sensory modality space (e.g. sounds or tactile sensations). Thus, defective eyes might be bypassed using a substituting sensory pathway. That this mapped or encoded information will reach the visual cortex and not elsewhere in the brain, is a fact rooted in the idea of natural brain plasticity. As a consequence, cortical re-mapping or reorganization happens when the brain is subject to either neural lesions or training. This latter training of course, turns out to be central to sensory substitution. In this thesis, we studied quite a number of cases that endorse such an idea. Furthermore, we considered clinical accounts that show activity in the visual cortexes of congenital blind individuals who underwent rehabilitation, using sensory substitution devices. In light of this, the present document explores among the most relevant sensory substitution devices, from Paul Bach-y-Rita's first attempts up to cutting-edge developments in this field. Then, the conclusion was drawn that

the most used modality to substitute vision is the auditory pathway. This is mostly the case owing the fact that the capability of the auditory sense to transmit information is the second greater in humans, only overtaken by that of the vision itself. Nonetheless, even though the ear is known to be capable of transmitting 10 Kbps, such capacity still lies far away from that of the vision, which may reach up to 1000 Kbps.

Looking back throughout the evolution of SSDs, it became clear that they all have tried to improve their sound outputs or sonic codes, as though ignoring the fact that no sound can lead to a full vision-like experience. In other words, they have focused on designing sensations, yet they do little for aiding the perceptual experience as such. It might be true that if an optimal audio-based sensation were to reach the visual area of the brain, the perception will occur naturally. Yet, building an optimal sensation out of sounds is not possible at all due to the large sensory information rate mismatch between vision and hearing. Notwithstanding, state-of-the-art SSDs keep on making more complex sounds such as soundscapes, in the hope to improve the perceptual experience. In theory it should improve (though never enough), yet in practice those sounds are bound to be confusing and even uncomfortable. In these instances, we argue that more needs to be done in order to enhance the perceptual experience of a SSD's user. By no means, however, the visual-to-sound encoding must be abandoned in SSDs. Quite the opposite; we promote additional, complementary and never exclusionary techniques to cope with the actual issue of the ears being unable to convey sufficient information as to create visual-like experiences. Our thesis is, hence, that the coding into sound of basic visual cues accompanied by computational methods that model higher perceptual levels of the visual system will lead us to a SSD: functional, ease to use, and suitable for mobility and exploration tasks. Such higher perceptual levels of vision cannot be better modeled by others than computer vision techniques.

Some would argue that the use of computer vision to recognize and then, communicate objects to the user triggers just visual imagery rather than actual vision via sensory substitution. For instance, letting a user know about the presence of a tree turns out to be general symbolic mapping mediated by the concept 'tree', whereas a soundscape may convey specific information of the scene: tree's type, perspective, location and so forth. However, we may add others [**14**], [**108**] who claim that "The only difference is that whereas imagining finds its information in memory, seeing finds it in the environment. Thus, one could say that vision is a form of imagining, augmented by the real world." As a consequence, 'normal' vision is itself constrained by top-down knowledge. This being known, it would be unpractical to deny to this knowledge a role in visual sensory substitution [**14**]. Top-down knowledge provides the kind of information that sighted individuals achieve from their visual systems, typically without conscious effort [**108**]. Furthermore, this computer-vision-based strategy will prevent the blind from spending 70 hours of training (and more) in recognizing an object that in any case, will never look as real as expected. In this order of ideas, we put forward thereafter in this document the concept of See ColOr.

What we have done

See ColOr whose name stands for *Seeing Colors with an Orchestra*, was introduced as a sensory substitution device that promotes context-awareness to the visually impaired and blind individual. In terms of hardware, the See ColOr prototype makes use of a 3D sensor (Microsoft Kinect), a light laptop, and a tactile tablet (iPad or iPhone). Our aim was to enlarge legibility of the nearby environment as well as to facilitate navigating towards desired locations, general exploration, and serendipitous discoveries. This document described in details the use of the audio and haptic trajectory playback to convey visual information that relates spatial awareness, revealing of objects, and obstacles perception. More specifically, the keystones studied in this document can be named as follow: research on sonification, enhancement of optical sensors, and research on haptic interfacing and computer vision. Overall, the SSD depicted in this work merges three levels of assistance: a global/local module for general exploration, an alerting method, and a recognition module:

I. An exploration module that makes it possible for the users to tap with their fingers the content of an image captured by a range camera in real time, while being rendered onto a tactile tablet. The color and position of explored points were mapped into sound by means of instruments and sound effects, respectively. This module exploits the audio and haptic trajectory feedback as a method to convey significant visual cues to the visually impaired. Particularly, the use of spatialized sound, which gives the illusion of virtual sources emitting from desired locations, served to emphasize spatial relations and shapes. It helped also to this purpose, the innate user's kinesthesia that maintains awareness of the fingers while sliding across the tablet.

II. An alerting method, based on range-imaging processing, that prevents the user from bumping into obstacles. It does so by informing about unexpected entities lying on the way and therefore, potentially leading to a fall. It was seen within this document as well that this algorithm could eventually predict the trajectory of encountered obstacles so as to maintain/suspend a warning, according to the likelihood of the collision. This method was introduced with the aim of helping the blind find a clear path in the interest of safe navigation.

III. A cognitive engine that uses state-of-the art object recognition methods to learn natural objects. It was presented as having two chief makings: an off-line training phase backed by detection/tracking methods, followed by a real-time searching process. This latter informs the user about the presence of previously learned objects during exploration, if any. Unlike I and II, this module perceives and depicts (through sounds) higher visual information out of which we derive sense of the world.

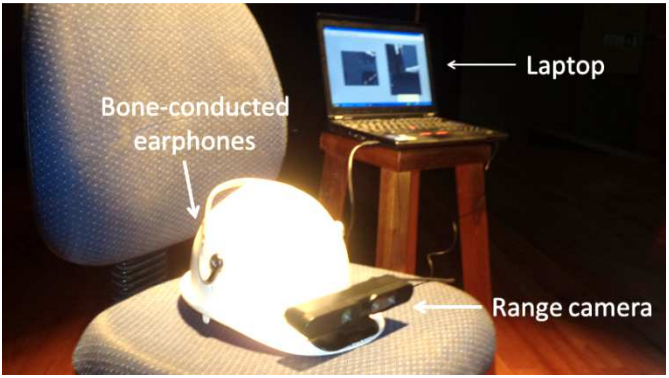


Figure 5-1. See ColOr prototype

The outcome of this research work was a prosthetic device that in contrast with related works, performs simultaneous coding of both, color and depth (in an RGB-D stream) into sound. Furthermore, this prototype makes use of computer vision methods for processing more complex visual features occasionally sonified as virtual objects/obstacles. Notice that the underlying software of this device was implemented in MATLAB Version 7.12.0.635 (R2011a) and even though real time was attained, drastically better performance is expected for a binary compiled version of core algorithms. By and large, for experiments reported in this thesis the results revealed that See ColOr is learnable, functional and provides easy interaction. Our encouraging results open a door towards autonomous mobility of the blind. More importantly though, these results served to confirm the central thesis of this work stating that: the coding into sound of basic visual cues (e.g. color and depth), accompanied by computational methods that model higher perceptual levels of the visual system (i.e. computer vision) may lead to more efficient SSD. Quite unlike many other devices in the state of the art, See ColOr proved to be a functional and utilitarian prototype (capable of functioning) that substitutes several features of vision at expense of relatively little user effort. In supporting this, several criteria were studied in this thesis, the most relevant one being the behavioral criterion. As a consequence, in moderate time, participants were enabled to grasp visual information of the world out of which they could derive: spatial awareness, ability to find someone, location of daily objects, and skill to walk safely avoiding obstacles.

Figure 5-2. Some frames of the video recording of a blind person finding someone.

Importantly, we need to put it clear enough that ours is still a research prototype of See ColOr that has been evaluated during lab experiments in highly controlled environments only. As a matter of fact, despite their promising potential and many years of development, SSDs have not been widely adopted yet. Only a few have ever been used outside of controlled research settings, and to the best of our knowledge no SSD whatsoever has been adopted as the main tool by a wide blind community [**33**]. In our opinion, this is because SSDs have reached already the limit defined by the physical mismatch (in terms of transmission of information) between vision and substituting senses. In light of this, we consider that our prototype by expanding the limits of typical SSDs towards the integration of computer vision, has made significant contributions to this field of research and showed interesting improvements compared to other SSD. Yet, a long way is still ahead before coming to a fully functional prosthetic device that can be used in a daily basis.

Reaching out end-users

Another key aspect of See ColOr is that the current prototype is made up of relatively affordable technologies. This is indeed beneficial for low-income countries, where blindness indexes continue to grow due to poor medical accessibility in rural areas. This also makes See ColOr more practical and better situated in terms of social impact. As proof of this, we had See ColOr tested in Colombia South America where the community involved in our experiments greatly enjoyed the experience of trying this sort of technology. Besides, out of this experience we collected important feedback that is bound to improve our system in the near future. For example, 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 the fact that besides being functional enough, an aid system must be wearable and comfortable. Particularly, we highlight the request of a non-negligible number of participants about relocating the camera. Over the years, blind individuals lose the instinctive notion of targeting his head 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. Likewise, they encouraged us to make further effort to spatialize sound also in height. Participants broadly showed enthusiasm and were acquiescent for the use of See ColOr. Those who had a guide-dog, however, expressed little interest in swapping it now, unless technology makes strides rapidly enough. This clearly leaves the door open for further research and stronger efforts in pursuit of more suitable prototypes.

As it was already mentioned, generally speaking, See ColOr enjoyed the acceptance and the appreciation of the blind and the visually impaired individuals who put it to the test. In this sense, a number of testimonial videos were recorded, some of which the readers can watch through the following links: video 1[18], video 2[19], video 3[20] and video 4[21]. Further, we

[18] https://www.youtube.com/watch?v=2rNTWTpu1-8
[19] https://www.youtube.com/watch?v=4309ojboYhk&noredirect=1

think it is worth highlighting some of the sentence that users were quoted saying: “suddenly all this amount of sonified colors becomes a symphony one has to master”; “one feels being a kind of musician and therefore one needs to further develop the auditory capabilities”; “This is simply something I have been always dreaming of”; “transmitting sound without covering my ears really rocks and helps”; “this system will definitely improve the quality of my life”; “I was born blind, so if you ask me, knowing the colors through sounds is an experience beyond my imagination”; “near an otherworldly experience”; “I did not know technology had reached this far”; “learning is not hard at all to me”. As for this latter comment, this work exposed the fact that using computer vision techniques in conjunction with speech prevents our users from spending 70 and more hours of training (e.g. the vOice). This is because the end result is a system absolutely intuitive that efficiently exploit user’s prior knowledge to avoid tough learning processes.

Figure 5-3. Some of our blind test-takers who agreed to be have pictures taken for this thesis.

Nevertheless, for lower level modules of See ColOr (i.e. global/local) learning of our color-to-sound coding is indeed requested. While mastering See ColOr sonic code will take several hours of training, this will be still negligible if we consider higher times for soundscape learning, let alone blind persons who may spend years on learning a braille system. Experiments reported in section Past Experiments involved a training phase to learn our color coding. For all our experiment participants, training lasted about 45 min. Afterwards, a small test for scoring the performance of the participants on sound/colors associations was achieved. On 15 heard sounds, the average number of correct colors among the six participants was 9.1 (standard deviation: 3.4). This clearly indicates that flawless association learning might take only several training sessions of the same duration. More specifically, one of our team members who masters See ColOr coding the best, claims to have invested a total of 30 hours before coming to a spontaneous association of colors and sounds.

More evidence in this regard is provided by Neil Harbisson, 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. He 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

[20] https://www.youtube.com/watch?v=FJyXftwwzks
[21] https://www.youtube.com/watch?v=QZ_t_BNWS7M

dream in colors". This is quite an important claim, as it means that his brain continues to create the sounds of the colors long after the camera has been shut down. Finally, even though learning of our coding seems to be reasonably low, we expect to further narrow it through the integration of our strategy for a relation between colors and sounds based on brain activity.

Otherwise, a number of concerns arise about having the users in See ColOr hold a tablet all the time. In this regard some say: "An issue is that the user's hands are always occupied by the iPad which reduces freedom of experiencing the space physically." And more importantly, others have said: "Is the idea that the blind user uses the screen limitations as an orientation guide as to their environments?" Actually, we cannot agree more with these critics, reason why we developed a solution to cope with this apparent issue. In this thesis, we introduced this strategy as "Tactile Augmented Reality", an idea whose primary implementation leaves little doubt about its convenience. Instead of gathering the coordinates of the fingers from the iPad screen to learn the point in the picture selected for sonification. We let the user enter his hands in the picture itself to point in the real world with his finger, so we track the fingertip and learn the point target for sonification. The novelty of this approach is that when a user points and object in the picture and touches it in the real world, the touching yields the sound of the object's color. In other words, he feels the natural sensations of touching the object's surface such as texture, temperature, resistance or elasticity; and in addition, he hears as well the color (instrument sound) and depth (rhythm) of the object. Therefore, his tactile sensation is augmented by color, a feature that has never been known to come from touch but sight. One other important aspect here is that Tactile Augmented Reality closes the perception-action loop, since users can now coordinate sensations with the 3D physical space. The video of our beta implementation (video[22]) further clears up this idea. Idea that certainly leads See ColOr to a quite a new experience on human machine interaction.

Importantly, Tactile Augmented Reality was not the only approach to radically different interaction methods that this thesis yielded. We also modeled, designed and fine-tuned a braille-like method for blind individuals to interact and understand natural text in the wild. Using deep neural networks and deep learning we turned our global module embedded in an iPhone, into a tactile system that reveals letters within the screen upon touch. Roughly, it works quite similar to the global module: images of the environment are captured, displayed in the iPhone screen and made touchable to the user. Differently though, touched points are no longer sonified as instruments sounds. Touched points in the image are sonified only if they belong to a letter or number and never otherwise. Exploring the screen carefully with their fingers, users are bound to discover all the alpha numeric components of the picture, having full access therefore to textual contents. Initial experiments finding buses numbers in the city of Geneva reveal the great potential of our approach and how it could be effectively used by blind users in a daily basis.

[22] https://www.youtube.com/watch?v=EWMpm_28Dlk

What we have learned

Throughout this research, many lessons were learned which are reflected in this thesis. Overall, in the one hand, we grew more sensitized on the importance of vision as a topic of research. Besides scientific challenges, we learned how humans are very visual animals that resulted seriously incapacitated when lacking this functionality. Hence, when it comes to sight handicapped individuals, no one doubts that every effort we put on improving their quality of life will be little. On the other hand, as computer scientists trained to tackle problems ruled by mathematics, the like of which we can model through variables. It became clear to us that humans are fairly more complex 'entities' ruled by feelings and emotions, which makes it harder to work with. Furthermore, the lesson learned is that this sort of research needs to extend beyond the laboratory, as it is always incumbent on a researcher to make the effort to reach out to the community to engage with the population he is serving. This is broadly known as community-based participatory research (CBPR): a research that has to be conducted as a close partnership between traditionally trained "experts" and members of a community, as otherwise, neither of them alone would suffice. Likewise, this work also left other more specific teachings out of which various research questions may be solved, namely:

- We learned that humans can indeed attain color and depth information accurately despite the lack of natural vision. It was shown through experiments in this thesis that the auditory pathway may be used as a substitute to this end. Importantly though, training is essential. Many blind and visually impaired individuals can benefit from this fact, since color provides clues for object identification and permits the communication with sighted individuals about the visual world, so they can share concepts on similar basis. In addition, depth is crucial to spatial awareness.
- It seems that nowadays, cutting-edge technology along with state-of-the-art computational methods, are capable to assist a blind person moderately well, when it comes to create a visual images through sounds. Nevertheless, if we look back to the 70s, and compare the visual images that Raymond M. Fish elicited in his patients using his earliest methods, to those images created using 3D-sound and computer vision. We will notice that small evolving steps in this line demand not years but, decades of technological advances. To make it worse, whether a day will come that an image can be elicited in detail in someone's mind by means of audio, is indeed a question that remains uncertain.
- Out of the available technology nowadays, we can engine systems that allow blind users behave somewhat like sighted individuals. We learned, however, that this is the case just for specific tasks and heavy constrained environments. We had, for instance, blind individuals in this work grasping objects and recognizing people accurately enough, as though they could see them. Yet, a very limited number of objects were used and reproducing these results outdoor is still challenging. Computational capacities still lag behind the needs for real time applications, whereas optical sensors need to grow lighter, cheaper and less restricted.

- Despite the increasingly popular use of speracons, auditory icons, earcons and even soundscapes. We learned that when it comes to communicate complex visual information such as, objects, faces and so forth. The use of natural speech is advisable at all, regardless its associated problems with language dependency or brain processing. This conclusion may be drawn from the experiments presented herein this work, but also owing the fact that it was a common belief among blind users.
- Quite opposite to our intuition, we learned in this work that the use of several fingers makes no difference for the blind people, when it comes to interact with touch screen. The use of one finger turns out to be more comfortable, more intuitive and never less useful. This could be a keystone for current designers of tactile interfaces aimed at promoting accessibility of the blind.

Future work

Our future research must flow towards the validation of See ColOr at the cognitive and neurological level. We stated in this thesis that our combination of sensory substitution and computer vision would lead us to a more efficient and functional SSD and, at the practical level, this proved to be right. Experiments reported in this thesis clearly show that our idea has stepped forward towards achieving autonomous mobility of the blind. Nevertheless, in the future, we would like to present evidence verifying that the addition of computer vision to sensory substitution might elicit enhanced visual consciousness. This is to say that the implications of our thesis reach beyond a functional prototype and lie straight inside the workings of the brain.

The central idea for future investigations is that computer vision triggers just visual imagery or memories that mediate the general concept an object, e.g. chair. This vague concept can be enhanced through the sonification of visual cues so as to achieve details of the specific chair, e.g. color, position, size, structure. The problem here is that such a primary or general concept is needed to make sense of the visual sounds. Without this initial concept the sounds are hardly coherent and, since there is no concept to attach them, they just drift away. This makes all sense since we saw through this thesis that conceptual information involved in vision cannot be conveyed through sound due to the bandwidth limit of the ear, among other reasons. Again, sensory substitution contributes with the sensations of the chair and computer vision does so with the perception as such.

A practical test to this hypothesis relies on the work of Gallant et al [**37**], who are able to reconstruct a mental image using fMRI-based brain scanning. The test could be done as follow: (a) Have a person imagine an object told by a computer vision system (CVS), e.g. chair. Then, reconstruct his mental image to see that even though the image is akin a chair, it has nothing to do with the actual chair the CVS is detecting. (b) Have a person use See ColOr to explore a chair placed in front, without letting him know it is a chair. Then, reconstruct his mental image to see that it is hardly akin to any chair. (c) Have the CVS inform the user

about the chair (general concept or perception) and let the user try See ColOr to explore the same chair at the same time (visual details or sensation). Then, reconstruct the mental image to see that unlike (a) and (b), this image is akin to the real chair being object of test.

Moreover, though the research presented in this thesis has answered many questions, some others have been raised too. For instance, there are practical issues arising from this work which should be pursued. Firstly, we could investigate thoroughly parallel or high-performance computing, with a view to increasing capability of our system for hardware-embedded or commercial prototypes. There are still modules we cannot afford to run in parallel (i.e. text recognition module with the rest) and more importantly, the object recognition task is nearing the real time limit. Thus, speeding up CPU times using multi-core computers will allow us to integrate more sophisticated methods for feature extraction and classification in object recognition. Particularly, we are interested in adding to See ColOr action/pose estimation and attribute classification commonly used in high-level computer vision[23]. For instance, poselets [**207**] can be used to recognize gender, hair style and types of clothes in natural scenes, e.g., *this person is male wearing glasses, jeans and t-shirt; he has long hair and no hat*. In this way, quite a number of relevant information could be made automatically accessible to unsighted individuals. Nevertheless, poselets are still fairly expensive [**207**] in computational terms, especially if intended in a multi-task (multi-module) embedded software [**207**]. To start with however, a full migration of our Matlab code into C++ is bound to speed up the system performance outstandingly.

A second line of practical research, which follows form chapter 5, would be to conduct more systematical experiments with end-users. Since the experimental evidence reported in this thesis precludes drawing any conclusions about differences (if at all) between blind and congenitally blind using See ColOr. As noted by [**129**] these populations must be studied separately since their cognition and interaction with the world are as different as theirs compared to ours (sighted people). Far more important yet is the conduction of outdoor experiments, which was not attempted at all in this thesis due to the limitations of current ToF cameras. To this end, our research must focus on stereo-vision technologies that meet both, accuracy and lightweight simultaneously. As a matter of fact, our lab is already advancing towards this topic in partnership with *Vision Embedded Systems, CSEM SA*[24]. A portable outdoors camera rig (the Icycam[25]) is being developed within the EyeWalker project [**235**]. We think that in the near future See ColOr could also benefit from the outcome of this research. Otherwise but still in the same line of testing with end-users, there is the need of experimentally proving two key ideas proposed in this thesis: the neural-based relation between color and sound (3.3.3 A relation between colors and sounds based on brain activity) and the concept of Tactile Augmented Reality (3.3.5 Tactile Augmented Reality: An alternative to the use of a tablet). These topics might certainly open two extensive lines of research. While the

[23] http://www.cs.berkeley.edu/~lbourdev/poselets/
[24] http://www.csem.ch/site/
[25] http://content.media.cebit.de/media/000079/0079824eng.pdf

former is thought to be densely enough for a PhD project, provided its premises are validated. The latter, in turn, will probably lead to new subareas of research in human computer interaction.

The end

Finally, as reflecting on the philosophical side of our work, we also grew very much interested in knowing how vision (a nonphysical phenomenon) comes into being, out of physical activity in a physical brain: why we draw a total blank on the nature of this transformation? More vividly, "*how is that the water of the brain becomes the wine of vision?*" to quote again English psychologist Nicholas Humphrey as he compares such a transition with a miracle. Likewise, let us raise a question that arguably fits better this work: are researchers on visual substitution bound to succeed in eliciting visual consciousness artificially, or not? Given the intricacy inherent to this question there would be little point pursuing an answer, were it not for all breakneck strides that we saw through this thesis: congenital blind people gaining brain activity in the visual cortex after electric tongue stimulation, or auditory inputs. Also, others have come to adopt sighted-like behaviors by means of mere skin stimulation, or audio-trajectory-playback. In the end, what all this means to us is that in the middle of so many uncertainties, our research is but a little fire lighting up the vast obscurity around us. An obscurity that makes us feel, from time to time, like trapped within the darkness of blindness.

Appendix A: List of publications

- J Gomez, G Bologna, and T Pun, "See ColOr (Seeing Colors with an Orchestra): A sensory substitution device for the visually impaired," ACM Transactions on Accessible Computing - TACCESS, p. submitted, 2013.

- J Gomez, G Bologna, and T Pun, "Efficient Registering of Color and Range Images," EURASIP Journal on Image and Video Processing - JIVP, vol. 26, no. 4, pp. 126-144, 2013.

- J Gomez, G Bologna , and T Pun, "A Visual World to be Heard: Touch and audio trajectory playback as non-visual cueing method to enlarge spatial awareness of unsighted individuals," Neurocomputing, p. submitted, 2013.

- G Bologna, J Gomez, and T Pun, "Vision Substitution Experiments with See ColOr," in 5th. International conference on the interplay between natural and Artificial computation IWINAC 2013, Palma de Mallorca - Spain, June, 2013.

- J Gomez, G Bologna, and T Pun, "Real-time Image Registration of RGB Webcams and Colorless 3D Time-of-flight Cameras," in International Conference on Computer Vision (ICCV), Florence - Italy, September, 2011.

- J Gomez, G Bologna, and T Pun, "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 CO - USA, October, 2012.

- J Gomez, G Bologna, and T Pun, "Spatial Awareness and Intelligibility for the Blind: Audio-touch Interfaces," in The ACM SIGCHI Conference on Human Factors in Computing Systems CHI, Austin TX - USA, March, 2012.

- J Gomez, G Bologna, and T Pun, "A virtual ceiling mounted depth-camera using orthographic kinect," in The International Conference on Computer Vision ICCV, Barcelona - Spain, June, 2011.

- J Gomez, S Mohammed, G Bologna, and T Pun, "Toward 3D Scene Understanding via Audio-description: Kinect-iPad fusion for the visually impaired," in International ACM SIGACCESS Conference on Computers and Accessibility, Dundee - UK, October, 2011.

- J Gomez, S Mohammed, and T Pun, "3D Scene accesibility for the blind via auditory-multitouch interfaces," in Open acccesibility everywhere: groundwork, infrastructure, standars AEGIS, Brussels - Belgium, November, 2011.

- J Gomez, G Bologna, and T Pun, "Multisource sonification for visual substitution in an auditory memory game: one, or two fingers," in The International Conference on Auditory Display ICAD, Budapest - Hungary, April, 2011.

- J Gomez, G Bologna, and T Pun, "Color-Audio encoding Interface for Visual Substitution: See ColOr Matlab-based Demo," in The 12th International ACM SIGACCESS Conference on Computers and Accessibility, Orlando, FL - USA, June, 2010.

- G Bologna, B Deville, J Gomez, and T Pun, "Toward local and global perception modules for vision substitution," Neurocomputing, vol. 74, no. 8, pp. 1182–1190, 2010.

Appendix B: Back-propagation rule deduction

This appendix corresponds to sub-section 3.6.5 Deep Neural Networks and Deep Learning

We start from the general equation of error in an artificial neural network:

$$E(w) = \frac{1}{2}\sum_{l=1}^{u}\sum_{i=1}^{s}(t_i^l - a_i^l)^2$$

where w represents the weights of the net; i sweeps all the u neurons in the output layer of the net; l sweeps all the s training patterns (i.e. as many inputs/outputs we want the net to learn); t_i^l is the expected or target output for the *l-th* pattern (input) at the *i-th* neuron; and a_i^l is the actual output for the *l-th* pattern (input) at the *i-th* neuron. A more general version of this error function can be written as follows:

$$E_{AV} = \frac{1}{N}\sum E(n)$$

with *E(n)* as the gradient of the function (net) over a single pattern in the *n-th* iteration of the training, i.e. instant gradient. Therefore, E_{AV} stands for the gradient over the whole set of patterns (N), i.e. the real gradient. The update of the weights using the real gradient is known as update by epochs. If the update is carried out by means of the instant gradient we call it sequential and its advantage is less computational load. The deduction of the rule for the gradient descent-based training based on the instant gradient is made so that:

$$w_{ji}(n+1) = w_{ji}(n) - \alpha\frac{\partial E(n)}{\partial w_{ji}(n)}$$

Here the calculation of the gradient depends on the place of the weight (w_{ji}) in the net (whether in the output layer or not). The calculation for a weight in the output layer can be deduced like this:

$$E(n) = \frac{1}{2}\sum_j e_j^2(n), \quad e_j(n) = d_j(n) - y_j(n), \quad y_j(n) = \phi_j(V_j(n)), \quad V_j(n) = \sum_i w_{ji}y_i(n)$$

So e is the error of the neuron j in the output layer; d_j is the target output of this neuron, and y_j is but its actual output. This actual output is given by V_j which is the activation function of the neuron. Applying the chain rule, we have:

$$\frac{\partial E(n)}{\partial w_{ji}(n)} = \frac{\partial E(n)}{\partial e_j(n)}\cdot\frac{\partial e_j(n)}{\partial y_j(n)}\cdot\frac{\partial y_j(n)}{\partial V_j(n)}\cdot\frac{\partial V_j(n)}{\partial w_{ji}(n)}$$

So that,

$$\frac{\partial E(n)}{\partial e_j(n)} = e_j(n), \frac{\partial e_j(n)}{\partial y_j(n)} = -1, \frac{\partial y_j(n)}{\partial V_j(n)} = \phi_j'(V_j(n)), \frac{\partial V_j(n)}{\partial w_{ji}(n)} = y_i(n)$$

$$\frac{\partial E(n)}{\partial e_j(n)} = -e_j(n)\phi_j'(V_j(n))\, y_i(n) \quad ,$$

therefore, $w_{ji}(n+1) = w_{ji}(n) - \alpha \frac{\partial E(n)}{\partial w_{ji}(n)} = w_{iji}(n) + \alpha\left[e_j(n)\phi_j'(V_j(n))\, y_i(n)\right]$,

for all the weights that arrive to the j neuron in the output layer we have:

$$w_{j1}(n+1) = w_{j1}(n) + \alpha e_j(n)\phi_j'(V_j(n))\, y_1(n)$$
$$w_{j2}(n+1) = w_{j2}(n) + \alpha e_j(n)\phi_j'(V_j(n))\, y_2(n)$$
.
.
$$w_{ji}(n+1) = w_{ji}(n) + \alpha e_j(n)\phi_j'(V_j(n))\, y_i(n)$$

Notice that the term $e_j(n)\phi_j'(V_j(n))$ is common to all the weights arriving to this neuron, so that we can define it as the gradient for the neuron *j*,

$$\delta_j = e_j(n)\phi_j'(V_j(n)).$$

Therefore, we are permitted to rewrite:

$$w_{j1}(n+1) = w_{j1}(n) + \alpha\delta_j y_1(n)$$
$$w_{j2}(n+1) = w_{j2}(n) + \alpha\delta_j y_2(n)$$
.
.
$$w_{ji}(n+1) = w_{ji}(n) + \alpha\delta_j y_i(n)$$

As a consequence, we can finally update any weight of the output layer as follows:

$$w_{ji}(n+1) = w_{ji}(n) - \alpha\frac{\partial E(n)}{\partial w_{ji}(n)} = w_{ji}(n) + \alpha\delta_j y_i(n) = w_{ji}(n) + \alpha e_j(n)\phi_j'(V_j(n))y_i(n),$$

for *j:1…s*, all the neurons in the output layer and *i:1…v*, all the neurons in the layer just before the output. Here α [0…1] is defined as the learning rate and it is the portion of the gradient that we want to follow. In other words, the gradient leads the algorithm towards to a minimal of the error function, but α regulates in which proportion or velocity we follow the direction given by the gradient. Big steps may take us beyond the minimal during the iterations, so α is typically chosen little. Now, to calculate the gradient of a weight in a hidden layer (before the output) we must proceed as follows:

.

. . .

. . $y_j(n)$. *output of the net*

$$\ldots \rightarrow \otimes^i \rightarrow \otimes^j \rightarrow \otimes^k \rightarrow \left[e_k(n) = d_k(n) - y_k(n)\right] \qquad (1)$$

. . .

. . .

.

$\frac{\partial E(n)}{\partial w_{ji}(n)} = \frac{\partial E(n)}{\partial V_j(n)} \cdot \frac{\partial V_j(n)}{\partial w_{ji}(n)}$, note that this relation is met for all the weights in the net.

Since $\frac{\partial V_j(n)}{\partial w_{ji}(n)} = y_i(n)$, *therefore,* $\frac{\partial E(n)}{\partial w_{ji}(n)} = \frac{\partial E(n)}{\partial V_j(n)} y_i(n)$

$\frac{\partial E(n)}{\partial w_{ji}(n)} = -\delta_j(n) y_i(n)$, where $\delta_j(n)$ is the local gradient in the neuron j of a hidden layer. This being said, the calculation of such a gradient must proceed like this:

$$\delta_j(n) = -\frac{\partial E(n)}{\partial V_j(n)}$$

$\frac{\partial E(n)}{\partial V_j(n)} = \frac{\partial E(n)}{\partial y_j(n)} \cdot \frac{\partial y_j(n)}{\partial V_j(n)}$, $\frac{\partial y_j(n)}{\partial V_j(n)} = \phi_j'(V_j(n))$

$$e_k(n) = d_k(n) - y_k(n) = d_k(n) - \phi(V_k(n)) = d_k(n) - \phi\left(w_{k1} y_1 + \ldots + w_{kj} y_j + \ldots\right)$$

Observing (1), we see that $y_j(n)$ is the output of the *j-th* neuron of a hidden layer and appears in all the errors of the neurons in the output layer, so that:

$$\frac{\partial E(n)}{\partial y_j(n)} = \frac{\partial\left[\frac{1}{2}\sum e^2(n)\right]}{\partial y_j(n)} = \sum_k e_k(n)\frac{\partial e_k(n)}{\partial y_j(n)}$$

$$\frac{\partial E(n)}{\partial y_j(n)} = \sum_k e_k(n)\frac{\partial e_k(n)}{\partial y_k(n)}.\frac{\partial y_k(n)}{\partial V_k(n)}.\frac{\partial V_k(n)}{\partial y_j(n)}$$

where...

$$\frac{\partial e_k(n)}{\partial y_k(n)} = -1, \frac{\partial y_k(n)}{\partial V_k(n)} = \phi_k^{'}(V_k(n)), \frac{\partial V_k(n)}{\partial y_j(n)} = w_{kj}$$

Therefore,

$$\frac{\partial E(n)}{\partial y_j(n)} = -\sum_k e_k(n)\phi_k^{'}(V_k(n))w_{kj}$$

Being $e_k(n)\phi_k^{'}(V_k(n)) = \delta_k$, the local gradient in the *k-th* neuron of the hidden layer, we have that $\frac{\partial E(n)}{\partial y_j(n)} = -\sum_k \delta_k(n)w_{kj}$.

Now we can retake the deduction of the rule as follows:

$$w_{ji}(n+1) = w_{ji}(n) - \alpha\left(\frac{\partial E(n)}{\partial w_{ji}(n)}\right)$$

$$\downarrow$$

$$\left(\frac{\partial E(n)}{\partial V_j(n)}.\frac{\partial V_j(n)}{\partial w_{ji}(n)}\right)$$

$$\downarrow \qquad \downarrow$$

$$\left[\frac{\partial E(n)}{\partial y_j(n)}.\frac{\partial y_j(n)}{\partial V_j(n)}\right] \quad [y_i(n)]$$

$$\downarrow \qquad \downarrow$$

$$-\sum_k \delta_k(n) w_{kj} \qquad \phi_j^{'}(V_j(n)) \cdot$$

Therefore,

$$\frac{\partial E(n)}{\partial w_{ji}(n)} = -y_i(n) \sum_k \delta_k(n) w_{kj} . \phi_j^{'}(V_j(n)) = -y_i(n) \phi_j^{'}(V_j(n)) \sum_k \delta_k(n) w_{kj} ,$$

where…

$$\delta_j(n) = \phi_j^{'}(V_j(n)) \sum_k \delta_k(n) w_{kj} = \phi_j^{'}(V_j(n))(\delta_1 w_{1j} + \delta_2 w_{2j} + ... + \delta_k w_{kj} + ... + \delta_m w_{mj})$$

$$\frac{\partial E(n)}{\partial w_{ji}(n)} = -y_i(n) \delta_j(n)$$

Finally the rule of that update weights in the ANN is given by:

$$w_{ji}(n+1) = w_{ji}(n) + \alpha y_i(n) \delta_j(n)$$

where $\phi_j^{'}(V_j(n))$ is the derivate of the activation function of the *j-th* neuron in the hidden layer evaluated in a local field. $\delta_k(n)$ is the local gradient of neurons in the output layer. And w_{kj} is the weight that links the *j-th* neuron in the hidden layer with the *k-th* neuron in the output layer. Note that the error is always being propagated backwards (previous layers). Therefore ANN trained with this rule are called back-propagation networks, since the tune the weights from output to input layer based on a gradient descend rule over an error function.

Appendix C: Author's biography

Juan D. Gomez attended Pereira's Tech. University of Colombia, where he majored –cum laude– in *Computer Science & Systems Engineering*. Right afterwards, he accomplished a year of research training in medical image processing at the ***MEM Institute for Surgical Technology and Biomechanics*** of the 'Universität Bern', in Switzerland. Later on, in Spain, he would receive two Master degrees: firstly, in *Computer-Vision & Artificial Intelligence* from the 'Universitat Autònoma' of Barcelona in 2009; and another in *Computer Science* from the 'Universidad Rey Juan Carlos' of Madrid in 2010. Both his master theses were conducted in parallel to his position as researcher at the *Computer Vision Center* –Barcelona– and the university hospital *Germans Trias i Pujol* –Badalona–. In the months to come, he will be awarded a Ph.D. in *Computer Science* by the 'Université de Genève' –Switzerland– for his research conducted on Visual Sensory Substitution within the *Computer Vision and Multimedia Lab*. While in his doctoral studies, Gomez performed his doctoral internship at the University of California Berkeley –US–, under supervision of American scientist Prof. Ruzena Bajcsy. Also, he had complementary formation in computer vision and machine learning at the ENS/INRIA institute Paris, France. His current research interests lie in the overlap of biological systems and computational models.

Bibliography

[1] world Health Organization. (2010, March) WHO. [Online]. http://www.who.int/mediacentre/factsheets/fs282/en/

[2] D Dorn, K Ahuja, and A Caspi, "The Detection of Motion by Blind Subjects with the Epiretinal 60-Electrode (Argus II) Retinal Prosthesis," *Archives of ophthalmology*, vol. 1, no. 1, pp. 1-7, 2012.

[3] E Zrenner, R Wilke, and H Sachs, "Patients allow recognition of letters and direction of thin stripes," in *World Congress on Medical Physics and Biomedical Engineering*, Munich - Germany, September, 2009.

[4] M Schmidmaier, "Sensory Substitution Systems," in *Media Informatics Advanced Seminar on Multimodal Human-Computer Interaction*, Santa Monica, CA - USA, June, 2011.

[5] P Maurice, M Solvej, G Albert, and K Ron, "Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind," *Brain*, vol. 128, no. 3, pp. 606–614, 2005.

[6] T Way and K Barner, "Automatic visual to tactile translation, part I: human factors," *IEEE Transactions on Rehabilitation Engineering*, vol. 1, no. 5, pp. 81-94, 1997.

[7] N Bradley and M Dunlop, "Investigating context-aware clues to assist navigation for visually impaired people," in *Workshop on Building Bridges, Interdisciplinary Context-Sensitive Computing*, Glasgow - UK, July, 2002.

[8] F Ribeiro, "Auditory augmented reality: Object sonification for the visually impaired," in *IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, Banff - Canada, September, 2012.

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

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

[11] P T Walling, "Consciousness: a brief review of the riddle," in *Annual Health Literacy Research Conference*, Washington - USA, May, 2000.

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

Press, 1998.

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

[14] J Ward and T Wright, "Sensory substitution as an artificially acquired synaesthesia," *Neuroscience and Biobehavioral Reviews*, vol. 1, no. 1, pp. 11–17, 2012.

[15] Christof Koch , *Consciousness: Confessions of a Romantic Reductionist*. Cambridge: MIT press, 2012.

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

[17] G Berkeley, *An Essay towards a New Theory of Vision*. New York: Aaron Rhames, 1709 (2008).

[18] J Gomez, G Bologna, and T Pun, "See ColOr (Seeing Colors with an Orchestra): A sensory substitution device for the visually impaired," *ACM Transactions on Accessible Computing - TACCESS*, p. submitted, 2013.

[19] J Gomez, G Bologna, and T Pun, "Efficient Registering of Color and Range Images," *EURASIP Journal on Image and Video Processing - JIVP*, vol. 26, no. 4, pp. 126-144, 2013.

[20] J Gomez, G Bologna , and T Pun, "A Visual World to be Heard: Touch and audio trajectory playback as non-visual cueing method to enlarge spatial awareness of unsighted individuals," *Neurocomputing*, p. submitted, 2013.

[21] G Bologna, J Gomez, and T Pun, "Vision Substitution Experiments with See ColOr," in *5th. International conference on the interplay between natural and Artificial computation IWINAC 2013*, Palma de Mallorca - Spain, June, 2013.

[22] J Gomez, G Bologna, and T Pun, "Real-time Image Registration of RGB Webcams and Colorless 3D Time-of-flight Cameras," in *International Conference on Computer Vision (ICCV)*, Florence - Italy, September, 2011.

[23] J Gomez, G Bologna, and T Pun, "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 CO - USA, October, 2012.

[24] J Gomez, G Bologna, and T Pun, "Spatial Awareness and Intelligibility for the Blind: Audio-touch Interfaces," in *The ACM SIGCHI Conference on Human Factors in Computing Systems CHI*, Austin TX -

USA, March, 2012.

[25] J Gomez, G Bologna, and T Pun, "A virtual ceiling mounted depth-camera using orthographic kinect," in *The International Conference on Computer Vision ICCV*, Barcelona - Spain, June, 2011.

[26] J Gomez, S Mohammed, G Bologna, and T Pun, "Toward 3D Scene Understanding via Audio-description: Kinect-iPad fusion for the visually impaired," in *International ACM SIGACCESS Conference on Computers and Accessibility*, Dundee - UK, October, 2011.

[27] J Gomez, S Mohammed, and T Pun, "3D Scene accesibility for the blind via auditory-multitouch interfaces," in *Open acccesibility everywhere: groundwork, infrastructure, standars AEGIS*, Brussels - Belgium, November, 2011.

[28] J Gomez, G Bologna, and T Pun, "Multisource sonification for visual substitution in an auditory memory game: one, or two fingers," in *The International Conference on Auditory Display ICAD*, Budapest - Hungary, April, 2011.

[29] J Gomez, G Bologna, and T Pun, "Color-Audio encoding Interface for Visual Substitution: See ColOr Matlab-based Demo," in *The 12th International ACM SIGACCESS Conference on Computers and Accessibility*, Orlando, FL - USA, June, 2010.

[30] G Bologna, B Deville, J Gomez, and T Pun, "Toward local and global perception modules for vision substitution," *Neurocomputing*, vol. 74, no. 8, pp. 1182–1190, 2010.

[31] P Bach-y-Rita, K Kaczmarek, E Tyler, and J Garcia-Lara, *Sustitucion sensorielle et qualia*. MIT Press, Boston: Boston, 1996.

[32] D Mary and A Burke. (2007, January) Emedicine. [Online]. http://emedicine.medscape.com/article/224309-overview

[33] Amir Amedi. (2009, June) Amir Amedi's Lab. [Online]. http://brain.huji.ac.il/site/em.html

[34] P Bach-y-Rita, "Seeing with the brain," *International Journal of Human Computer Interaction*, vol. 15, no. 2, pp. 285–295, 2003.

[35] K J O'Regan, *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*. Oxford: Oxford University Press, 2011.

[36] Julia Chai Jean, *Gian Paolo Lomazzo and the art of expression*. Cambridge: Harvard University, 1990.

[37] N Thomas, J Ryan, N Kendrick, and J Gallant, "Bayesian Reconstruction of Natural Images from Human Brain Activity," *Neuron*, vol. 63, no. 1, pp. 902–915, 2009.

[38] P Bach-y-Rita and W Kercel., "Sensory substitution and the humanmachine," *Trends in Cognitive Sciences*, vol. 7, no. 12, pp. 541–546, 2003.

[39] N Humphrey , *A History of the Mind: Evolution and the Birth of Consciousness*. New York: Springer, 1999.

[40] Oliver Sacks, *The Man Who Mistook His Wife for a Hat*. New York: Simon & Schuster, 1985.

[41] D Ghose and M Wallace, "Impact of response duration on multisensory integration," *Neurophysiol*, vol. 108, no. 9, pp. 2534-2544, 2012.

[42] N Aaron. (2010, January) Multisensory Research Laboratory at Vanderbilt University. [Online]. http://www.kc.vanderbilt.edu/multisensory/index.html

[43] H McGurk and J MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 1, pp. 746-748, 1976.

[44] M Micah and T Mark, *The Neural Bases of Multisensory Processes (Frontiers in Neuroscience)*. Boca Raton: CRC Press, 2011.

[45] D Felleman and D Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cereb Cortex*, vol. 1, no. 1, pp. 1– 47, 1991.

[46] M Abeles, *Corticonics*. Cambridge: Cambridge UP, 1991.

[47] M Young, K Tanaka, and S Yamane, "On oscillating neuronal responses in the visual cortex of the monkey," *Neurophysiol*, vol. 64, no. 1, pp. 1464 –1474, 1992.

[48] W Singer, "Synchronization of cortical activity and its putative role in information processing and learning," *Annu Rev Physiol*, vol. 55, no. 1, pp. 349 –374, 1993.

[49] W Singer, "Visual feature integration and the temporal correlation hypothesis," *Annu Rev Neurosci*, vol. 18, no. 1, pp. 555–586, 1995.

[50] N Hadjikhani1 and R Per, "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, pp. 1072–1084, 1998.

[51] N Logeswarana and B Joydeep, "Crossmodal transfer of emotion by music," *Neuroscience Letters* , vol. 455 , no. 2, pp. 129–133, 2009.

[52] B Meier and N Rothen, "Training grapheme-colour associations produces a synaesthetic Stroop effect, but not a conditioned synaesthetic response," *Neuropsychologia*, vol. 47 , no. 4, pp. 1208–1211, 2009.

[53] S Levy-Tzedek et al., "Cross-sensory transfer of sensory-motor information: visuomotor learning affects performance on an audiomotor task, using sensory-substitution," *Scientific Reports* , vol. 2, no. 1, pp. 949-960, 2012.

[54] N Tal and A Amedi, "Multisensory visual–tactile object related network in humans: insights gained using a novel crossmodal adaptation approach," *Exp Brain Res*, vol. 2, no. 3, pp. 165-82, 2009.

[55] Board of Regents of the University of Wisconsin System. (2012, August) Mitchell E. Tyler. [Online]. https://directory.engr.wisc.edu/bme/faculty/tyler_mitchell

[56] P Bach-y-Rita, K Kaczmarek, M Tyler, and M Garcia-Lara, "From perception with a 49-point electrotactile stimulus array on the tongue: a technical note," *Journal of Rehabilitation Research and Development*, vol. 35, no. 1, pp. 427–430, 1998.

[57] R F Wang and E S Spelke, "Human spatial representation: insights from animals," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 376-382, 2000.

[58] M Loomis, G Golledge, L Klatzky, J Speigle, and J Tietz, "Personal guidance system for the visually impaired," in *ACM SIGCAPH Computers and the Physically Handicapped*, New York - USA, March, 1994.

[59] E C Tolman, "Cognitive maps in rats and men," *Psychol*, vol. 55, no. 4, pp. 189-208, 1948.

[60] M Downs and D Stea, "Cognitive maps and spatial behaviour: process and products," *Image and Environment*, vol. 6, no. 2, pp. 8–26, 1997.

[61] M Ptito and S Desgent, *Sensory input-based adaptation and brain architecture*. Cambridge: Cambridge University Press, 2004.

[62] M Kitchin, M Blades, and R Golledge, "Understanding spatial concepts at the geographic scale

without the use of vision," *In Personal Ubiquitous Comput*, vol. 9, no. 6, pp. 395-403, 1997.

[63] R C Galistel, *The organization of learning*. Cambridge: MA : MIT Press, 1990.

[64] Miyoshi Takei. (2013, January) Tennis for the blind, Japan-style! [Online]. http://japandailypress.com/tag/miyoshi-takei

[65] B Han, C Paulson, J Wang, and D Wu, "Depth-Based Image Registration," in *Algorithms for Synthetic Aperture Radar Imagery*, New York - USA, January, 2010.

[66] Z Zhengyou, "Microsoft Kinect Sensor and Its Effect," *MultiMedia, IEEE* , vol. 19, no. 2, pp. 4 - 10, 2012.

[67] Department of Electrical and Computer Engineering : University of California Davis. (2011, February) The CIPIC Interface Laboratory Home Page. [Online]. http://interface.cipic.ucdavis.edu/

[68] B Deville, G Bologna, M Vinckenbosch, and T Pun, "Depth-based detection of salient moving objects in sonified videos for blind users," in *International Conference on Computer Vision Theory and Applications*, Funchal - Portugal, January, 2008.

[69] M Bujacz , "Representing 3D scenes through spatial audio in an electronic travel aid for the blind," Technical University of Lodz, Łódź, PhD thesis 0978, 2010.

[70] R Laurent et al., "Cross-modal activation of visual cortex during depth perception using auditory substitution of vision," *NeuroImage*, vol. 26, no. 1, pp. 573– 580, 2005.

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

[72] Y Visell, "Tactile sensory substitution: Models for enaction in HCI," *Interacting with Computers*, vol. 21, no. 1, pp. 38–53, 2009.

[73] T Marcinkowski, "Doktoraty HC: Prof. Witold Starkiewicz," *Medyk-Czasopismo lekarzy i studentów*, vol. 1, no. 1, pp. 10-12, 1991.

[74] P Bach-y Rita, C Collins, F Saunders, and B White, "Vision substitution by tactile image projection," *Nature*, vol. 221, no. 1, pp. 963–964, 1969.

[75] P Bach-y-Rita, "Tactile sensory substitution studies," *Annals of New York Academic Sciences*, vol. 1013, no. 1, pp. 83–91., 2004.

[76] P Bach-y-Rita, *Brain Mechanisms in Sensory Substitution*. New York: Academic Press, 1972.

[77] E Jhon Harrison and Baron-Cohen Simon , *Synaesthesia: classic and contemporary readings*. Oxford: Oxford : Blackwell, 1996.

[78] P Meijer , "An Experimental System for Auditory Image," *IEEE Transactions on Biomedical Engeneering*, vol. 39, no. 2, pp. 112-121, 1992.

[79] P Arno, C Capelle, M-C Wanet-Defalque, H Catalan-Ahumada, and C Veraat, "Auditory coding of visual patterns for the blind," *Perception*, vol. 28, no. 1, pp. 1013-1029, 1999.

[80] S Hanneton, "The vibe : a versatile vision-to audition sensory substitution device," *Applied Bionics and Biomechanics*, vol. 7, no. 4, pp. 269-276, 2010.

[81] J Bliss, M Katcher, C Rogers, and R Shepard, "Optical-to-tactile imageconversion for the blind," *IEEE Transactions on Man–Machine Systems*, vol. 11, no. 1, pp. 58–65, 1970.

[82] S Wall and S Brewster, "Sensory substitution using tactile pin arrays: human factors, technology, and applications," *Signal Processing*, vol. 86, no. 1, pp. 3674–3695, 2006.

[83] T Pun, "Simplification automatique de scenes par traitement numerique d'images en vue d'une restitution tactile pour handicapes de la vue," EPFL, Lausanne - Switzerland, PhD Thesis 425, 1982.

[84] T Pun, "Tactile Artificial Sight: Segmentation of Images for Scene Simplification," *IEEE Transactions on Biomedical Engineering*, vol. 29, no. 4, pp. 293-299, 1989.

[85] T Pun, "A new method for grey-level picture thresholding using the entropy of the histogram," *Signal Processing*, vol. 2, no. 3, pp. 223–237, 1980.

[86] H Kajimoto, M Inami, N Kawakami, and S Tachi, "Smarttouch: electric skin to touch the untouchable," *IEEE Computer Graphics and Applications*, vol. 24, no. 1, pp. 36–43, 2004.

[87] M Raymond, "An Audio Display for the Blind," *IEEE Transactions on Biomedical Engineering*, vol. 23, no. 2, pp. 144-154 , 1975.

[88] T Bower, "Blind babies see with the ears," *newScientist*, pp. 255-260, February 1977.

[89] L Kay, "Ultrasonic spectacles for the blind," in *Conf Sensory Devices for the Blind*, London, 1966.

[90] D Heyes, "Human navigation by sound," *Physics in Technology*, vol. 14, no. 2, pp. 68-75, 1983.

[91] L Merabet, L Battelli, S Obretenova, and S Maguire, "Functional recruitment of visual cortex for sound encoded object identification in the blind," *Neuroreport*, vol. 20, no. 1, pp. 132–138, 2009.

[92] C Capelle and C Trullemans, "A Real-Time Experimental Prototype for Enhancement of Vision Rehabilitation Using Auditory Substitution," *IEEE Transactions on Biomedical Engeneering*, vol. 45, no. 10, pp. 1279-1293, 1998.

[93] L Gonzalez-Mora, A Rodriguez-Hernandez, and N Sosa, "Development of a new space perception system for blind people, based on the creation of a virtual acoustic space," *Lecture Notes in Computer Science*, vol. 1607, no. 1, pp. 321–330, 1999.

[94] K Doel, "Sound view: sensing color images by kinesthetic audio," in *International Conferenceon Auditory Display (ICAD)*, Boston, MA - USA, July, 2003.

[95] D Barthélémy, L Nicolas, D Alleysson, and J Hérault, "Visuo-auditory sensory substitution for mobility assistance: testing TheVIBE," in *Workshop on Computer Vision Applications for the Visually Impaired*, Paris - France, October, 2008.

[96] Z Capalbo and B Glenney, "Hearing color: radical plurastic realism and SSDs," in *the Fifth Asia-Pacific Computing and Philosophy Conference*, Tokyo - Japan, October, 2009.

[97] S Levy-Tzedek et al., "Cross-sensory transfer of sensory-motor information: visuomotor learning affects performance on an audiomotor task, using sensory-substitution," *Scientific Reports*, vol. 2, no. 1, pp. 882–890, 2012.

[98] L Kai, L Tak , C Chi , and L Yunhui, "A Wearable Stereo Vision System for Visually Impaired," in *International Conference on Mechatronics and Automation*, Chengdu - China, June, 2012.

[99] T Winlock, E Christiansen, and S Belongie, "Toward Real-Time Grocery Detection for the Visually Impaired," in *Workshop on Computer Vision Applications for the Visually Impaired*, San Francisco, CA - USA, June, 2009.

[100] D G Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Journal Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[101] B Katz, P Truillet, S Thorpe, and C Jouffrais, "NAVIG: Navigation Assisted by Artificial Vision and GNSS," in *Workshop on Multimodal Location Based Techniques for Extreme Navigation*, Helsinki - Finland, May, 2010.

[102] S Kammouna et al., "Navigation and space perception assistance for the visually impaired: The NAVIG project," *IRBM biomedical engineering research journal*, vol. 33, no. 2, pp. 182–189, 2012.

[103] Christophe Jouffrais. (2008, November) NAVIG - Navigation Assistée par VIsion embarquée et GNSS. [Online]. http://navig.irit.fr/

[104] M Bujacz, P Skulimowski, and P Strumiłło, "Sonification of 3D scenes using personalized spatial audio to aid visually impaired persons," in *International Conference in Auditory Display ICAD*, Budapest - Hungary, June, 2011.

[105] Blue Water Media. (2012, January) Foundation Fighting Blindness. [Online]. http://www.blindness.org/

[106] F David , K Madeleine, and T Frank , "Haptic Guidance: Experimental Evaluation of a Haptic Training," in *10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, Chicago, IL - USA, March, 2002.

[107] F Dramas, S Thorpe, and S Jouffrais, "Artificial Vision For The Blind: A Bio-Inspired Algorithm For Objects And Obstacles Detection," *International Journal of Image and Graphics, World Scientific*, vol. 10, no. 4, pp. 531-544, 2010.

[108] the Andrea Bocelli Foundation. (2012, January) MIT Fifth Sense Project: Providing the Functions of Sight to Blind People. [Online]. http://people.csail.mit.edu/seth/misc/bocelli.html

[109] M Hersh and M Johnson, *Assistive technology for visually impaired and blind people*. New York: Spinger, 2008.

[110] K M Hou, E Pissaloux, H L Shi, K Ramli, and D Sudiana, "SEES: Concept and Design of a Smart Environment Explorer Stick," in *Human System Interaction International Conference HSI*, Gdansk - Poland, June, 2013.

[111] L Z Zhang et al., "Smart Environment Explorer Stick (SEES): Concept and Design of Its Orientation and Navigation System," in *International Workshop of NSICST 'New and Smart Information Communication Science and Technology'*, Clermont Ferrand - France, September, 2013.

[112] M Yusro et al., "Design and Implementation of SEE-Phone in SEES (Smart Environment Explorer Stick)," in *International Workshop of NSICST 'New and Smart Information Communication Science and Technology'*, Clermont Ferrand - France, September, 2013.

[113] M Benjamin, N Ali, and A Schepis, "A laser cane for blinds," in *San Diego Biomedical Symposium*, San Diego, CA - USA, June, 1973.

[114] J Borenstein and I Ulrich, "The guide cane, a computerized travel aid for the active guidance of blind pedestrians," in *IEEE International Conference on Robotics and Automation*, Albuquerque, CA - USA, June, 1997.

[115] B Hoyle and S Dodds, "The Ultra Cane mobility aid at work," in *Conference on Visual and Hearing Impairments*, Granada - Spain, January, 2006.

[116] R Farcy, "Une aideé lectronique miniature pour les dé placements des déficients visuels en intérieur," in *the Handicap*, Paris - France, March, 2008.

[117] Petrie, H, "User requirements for a GPS-based travel aid for blind people," in *Conference on Orientation and Navigation Systems for Blind people*, New York - USA, June, 1995.

[118] H Petrie, V Johnson, T Strothotte, A Raab, and S Fritz, "MoBIC: designing a travel aid for blind and elderly people," *Navigat*, vol. 49, no. 1, pp. 45-52, 1996.

[119] C M LaPierre, "Navigation System for the Visually Impaired," Canada, 2003.

[120] J Loomis, J Marston, R Golledge, and R Klatzky, "Personal guidance system for visually impaired people," *Visual Impairment Blindness*, vol. 99, no. 4, pp. 219–232, 2005.

[121] P Ponchilla, E Rak, A Freeland, and J LaGrow, "Accessible GPS: reorientation and target location among users with visual impairments," *Visual Impairment Blindness*, vol. 101, no. 7, pp. 389–401, 2007.

[122] D Rodriguez-Losada, F Matia, and R Galan, "Building geometric feature based maps for indoor service robots," *Robot. Autonomous Systems*, vol. 54, no. 7, pp. 546–558, 2006.

[123] J R Marston, "Towards an accessible city: empirical measurement and developing of access to urban opportunities for those with vision impairments," University of California, Santa Barbara, PhD thesis 00212, 2002.

[124] L Ran, S Helal, and M Drishti, "An integrated indoor/outdoor blind navigation system and service," in *the Second IEEE International Conference on Pervasive Computing and Communications*, New York - USA, December, 2004.

[125] J C Bliss, "Reading machines for the blind," in *Active Touch: The Mechanism of Recognition of Objects by Manipulation*. Oxford: Pergamon Press, Oxford, 1978, pp. 100-129.

[126] J Pasquero and V Hayward, "STRESS: a practical tactile display systems with one millimeter spatial resolution and 7000 Hz refresh rate," in *Eurohaptics*, Dublin - UK, June, 2003.

[127] R Velázquez, E Pissaloux, M Hafez, and J Szewczyk, "Tactile rendering with shape memory alloy pin-matrix," *IEEE Trans. Instrumentat. Meas.*, vol. 57, no. 5, pp. 1051–1057, 2008.

[128] C Wagner, S Lederman, and R Howe, "A tactile shape display using RC servomotors," in *10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, Orlando, FL - USA, March, 2002.

[129] A Crossan and S Brewster, "Multimodal Trajectory Playback for Teaching Shape Information and Trajectories to Visually Impaired Computer Users," *ACM Transactions on Accessible Computing*, vol. 1, no. 2, pp. 120-125, 2008.

[130] L Brown and S Brewester, "Drawing by ear: Interpreting sonified line graphs," in *The International Conference on Auditory Display*, Boston, MA - USA, June, 2003.

[131] L Alty and D Rigas, "Communicating graphical information to blind users using music: The role of context," in *Conference on Human Factors in Computing Systems*, Los Angeles, CA - USA, March, 1998.

[132] H Zhao, C Plaisant, and B Sheneiderman, ""I hear the pattern": Interactive sonification of geographical data patterns," in *the ACM SIGCHI Conference on Human Factors in Computing Systems*, Portland, OR - USA, March, 2005.

[133] M Kamel, P Roth, and R Sinha, "Graphics and user's exploration via simple sonics (GUESS): Providing interrelational representation of objects in a non-visual environmen," in *The International Conference on Auditory Display*, Espoo - Sweden, August, 2001.

[134] E Pissaloux, "Neuro-cognitive approach to visually impaired mobility and ICT supports," in *SIGNaL PROCESSING, algorithms, architectures, arrangements, and applications*, Poznan - Poland, September, 2013.

[135] V Goffaux, C Jacques, A Mouraux, and A Oliva, "Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence," *Visual Cognition*, vol. 12, no. 6, pp. 878–892, 2005.

[136] G Rousselet, O Joubert, and M Fabre-Thorpe, "How long to get to the "gist" of real-world natural scenes?," *Visual cognition*, vol. 12, no. 6, pp. 852-877, 2005.

[137] A Torralba, "How many pixels make an image?," *Visual Neuroscience*, vol. 26, no. 1, pp. 123–131, 2009.

[138] J Rossi, F Perales, J Varona, and M Roca, "Col.diesis: transforming colour into melody and implementing the result in a colour sensor device," in *The Second International Conference in Visualisation*, Barcelona - Spain, June, 2009.

[139] S Meers and K Ward, "A vision system for providing 3D perception of the environment viatranscutaneous electro-neura lstimulation," in *the Eighth International Conference on Information Visualisation*, Washington - USA, June, 2004.

[140] S Meers and K Ward, "A vision system for providing the blind with 3d colour perception of the enviroment," in *the Asia-Pacific Workshop on Visual Information Processing*, Tokyo - Japan, March, 2005.

[141] C Sjöström and K Rassmus-Gröhn, "The sense of touch provides new computer interaction techniques for disabled people," *Technology and Disability*, vol. 10, no. 1, pp. 45-52, 2010.

[142] L Cappelletti, M Ferri, and G Nicoletti, "Vibrotactile colour rendering for the visually impaired within the VIDET project," *Telemanipulator and Telepresence Technologies*, vol. 35, no. 24, pp. 92–96, 1999.

[143] J Tapson et al., "The Feeling of Color: A Haptic Feedback Device for the Visually Disabled," in *IEEE Biomedical Circuits and Systems Conference*, New York, 2008.

[144] N Hadjikhani and P Roland, "Cross-modal transfer of information between the tactile and the visual representations in the human brain: A positron emission tomographic study," *Journal of Neuroscience*, vol. 18, no. 3, pp. 77-84, 1998.

[145] D Nikolić, "Is synaesthesia actually ideaestesia? An inquiry into the nature of the phenomenon," in *International Congress on Synesthesia, Science & Art*, Granada - Spain, June, 2009.

[146] N Logeswaran and J Bhattacharya, "Crossmodal transfer of emotion by music," *Neuroscience Letters*, vol. 44, no. 2, pp. 129-133, 2009.

[147] R Davidson, G Schwartz, C Saron, J Bennett, and D Goleman, "Frontal versus parietal EEG asymmetry during positive and negative affect," *Psychophysiology*, vol. 16, no. 1, pp. 202-206, 1997.

[148] T Canli, J Desmond, Z Zhao, G Glover, and J Gabrieli , "Hemisphericasymmetry for emotional stimuli detected with fMRI," *Neuroreport*, vol. 9, no. 1, pp. 3233-3239, 1998.

[149] BioSemi products. (2000, January) BioSemi. [Online]. http://www.biosemi.com/

[150] P Zhang, "Neural Networks for Classification: A Survey," *IEEE Transactions on systems, man, and cybernetics*, vol. 30, no. 4, pp. 451-461, 2000.

[151] R Begault, *3-D Sound for Virtual Reality and Multimedia*. Boston: AP Professional, 1994.

[152] C Brown and R Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 100-105, 1998.

[153] A Dodson, T Moore, and G Moon, "A Navigation System for the Blind Pedestrian," in *3rd European Symposium on Global Navigation Satellite Systems*, Genova - Italy, March, 1999.

[154] N Franklin, "Language as a means of constructing and conveying cognitive maps," in *The construction of cognitive maps*. Dordrecht: Kluwer Academic Publishers, 1995, pp. 100-125.

[155] K Hemenway, "Psychological issues in the use of icons in command menus," in *The Conference on Human Factors in Computer Systems (CHI)*, New York - USA, August, 1982.

[156] P. Kolers, "Some formal characteristics of pictograms," *American Scientist*, vol. 57, no. 1, pp. 348-363, 1969.

[157] W Gaver, "Auditory icons: Using sound in computer interfaces," *Human-Computer Interaction*, vol. 2, no. 1, pp. 167-177, 1986.

[158] M Blattner, A Sumikawa, and R Greenberg, "Earcons and icons: Their structure and common design principles," *Human-Computer Interaction*, vol. 4, no. 1, pp. 11-44, 1989.

[159] B Walker, A Nance, and J Lindsay, "Spearcons: Speech-based earcons improve navigation performance in auditory menus," in *Conference on Auditory Display (ICAD)*, London - UK, June, 2006.

[160] B Huhle, P Jenke, and W Strasser, "On-the-fly scene acquisition with a handy multi-sensor system," *International Journal of Intelligent Systems Technologies and Applications*, vol. 5, no. 3, pp. 255-263, 2008.

[161] D Scaramuzza, A Harati, and R Siegwart, "Extrinsic self calibration of a camera and a 3D laser range finder from natural scenes," in *International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA - USA, June, 2007.

[162] D Herrera, J Kannala, and J Heikkila, "Joint depth and color camera calibration with distortion correction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2058-2064, 2012.

[163] M Van den Bergh and L Van Gool, "Combining RGB and ToF Cameras for Real-time 3D Hand Gesture Interaction," in *IEEE Workshop on Applications of Computer Vision (WACV)*, Cancún - Mexico, May, 2011.

[164] S Fuchs and G Hirzinger, "Extrinsic and depth calibration of ToF cameras," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage - Alaska, June, 2008.

[165] J Zhu, L Wang, R Yang, and J Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage - Alaska, June, 2008.

[166] B Zitová and J Flusser, "Registration methods: a survey," *Elsevier. Image and Vision Computing*, vol. 21, no. 11, pp. 977-1000, 2003.

[167] M Deshmukh and U Bhosle, "A survey of image registration," *International Journal of Image Processing (IJIP)*, vol. 5, no. 3, pp. 245-269, 1992.

[168] S H Yang, "Neural network based stereo matching algorithm utilizing vertical disparity," in *Annual Conference on IEEE Industrial Electronics Society (IECON)*, Phoenix, AZ - USA, May, 2010.

[169] J Sun, N Zheng, and H Shum, "Stereo Matching Using Belief Propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 787-800, 2002.

[170] P Biber and W Straßer, "The Normal Distributions Transform: A New Approach to Laser Scan Matching," in *International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV - USA, June, 2003.

[171] H Golub and L Van, *third ed. John Hopkins University*. Baltimore: Johns Hopkins University Press, 1996.

[172] A Staranowicz and G L Mariottini, "A Comparative Study of Calibration Methods for Kinect-style cameras," in *International Conference on Pervasive Technologies Related to Assistive Environments (PETRA)*, Crete - Greece, January, 2012.

[173] C Zhang and Z Zhang, "Calibration between Depth and Color Sensors for Commodity Depth Cameras," in *Workshop on Hot Topics in 3D (ICME)*, San Jose, CA - USA, June, 2011.

[174] D Herrera, J Kannala, and J Heikkila, "Accurate and Practical Calibration of a Depth and Color Camera Pair," in *14th International Conference in Computer Analisys Images (CAIP)*, Seville - Spain, December, 2011.

[175] A Staranowicz, F Morbidi, and G L Mariottini, "Depth-camera calibration toolbox (dcct): accurate, robust, and practical calibration of depth cameras," in *British Machine Vision Conference (BMVC)*, Guildford - UK, May, 2012.

[176] B Han, C Paulson, and D Wu, "Depth Based Image Registration via 3D Geometric Segmentation," *Journal of Visual Communication and Image Representation (JVCIR)*, vol. 22, no. 5, pp. 421-431, 2012.

[177] D Crispell, J Mundy, and G Taubin, "Parallax-Free Registration of Aerial Video," in *British Machine Vision Conference (BMVC)*, Leeds - UK, May, 2008.

[178] D Marr and T Poggio, "A computational theory of human stereo vision," *Theory of Light Absorption and Non-Radiative Transitions in F-Centres*, vol. 204, no. 1156, pp. 301-28, 1979.

[179] M Andersen et al., "Kinect Depth Sensor Evaluation for Computer Vision Applications," Aarhus, 2012.

[180] Dutta Tilak Dutta, "Evaluation of the kinect sensor for 3-d kinematic measurement in the workplace," *Applied Ergonomics*, vol. 43, no. 4, pp. 645-649, 2011.

[181] B Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 17-32, 1981.

[182] E Oja, "Neural networks, principal components, and subspaces," *International Journal of Neural Systems*, vol. 1, no. 1, p. 4871, 1989.

[183] F L Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567-585, 1989.

[184] Bouguet Jean-Yves. (2010, July) Camera Calibration Toolbox for Matlab. [Online]. http://www.vision.caltech.edu/bouguetj/calib_doc/

[185] Z Guo, "Reduced complexity Schnorr-Euchner decoding algorithms for MIMO systems," *IEE Communication Letters*, vol. 8, no. 5, pp. 286-288, 2004.

[186] J Anderson and S Mohan, "Sequential coding algorithms: A survey and cost analysis," *IEEE Transactions on Communications*, vol. 32, no. 6, pp. 169-176, 1994.

[187] M Ronald, G Jonathan, A William, and G Saul, *Readings in human–computer interaction. Toward the Year 2000*. San Francisco: Morgan Kaufmann, 1995.

[188] R Yang, S Park, S Mishra, C Newsom, and H Joo, "Supporting spatial awareness and independent wayfinding for pedestrians with visual impairments," in *The 13th International ACM SIGACCESS Conference on Computers and Accessibility*, Dundee - UK, September, 2011.

[189] S Kane, J Wobbrok, and R Lander, "Usable Gestures for Blind People: Understanding Preference and Performance," in *The annual conference on Human factors in computing systems*, Vancouver - Canada, June, 2011.

[190] D McGookin, S Brewster, and W Jiang, "Investigating touchscreen accessibility for people with visual impairments," in *The main Nordic forum for human-computer interaction research*, Lund - Sweden, March, 2008.

[191] M Kaltenbrunner, T Bovermann, R Bencina, and E Costanza, "TUIO - A Protocol for Table-Top Tangible User Interfaces," in *the 6th International Workshop on Gesture in Human-Computer Interaction and Simulation*, Vannes - France, May, 2005.

[192] M Kaltenbrunner and R Bencina, "reacTIVision: A Computer-Vision Framework for Table-Based Tangible Interaction," in *first international conference on "Tangible and Embedded Interaction" (TEI07)*, Baton Rouge - france, June, 2007.

[193] M Wright, A Freed, and A Momeni, "OpenSound Control: State of the Art 2003," in *3rd Conference on New Instruments for Musical Expression (NIME 03)*, Montreal - Canada, June, 2003.

[194] Martin Kaltenbrunner. (2010, December) TUIO protocol. [Online]. http://www.tuio.org/

[195] A Crossan and S Brewster, "Two-Handed Navigation in a Haptic Virtual Environment," in *The ACM SIGCHI Conference on Human Factors in Computing Systems CHI*, Montreal - Canada, May, 2006.

[196] C Ordonez, G Navarun, and A Barreto, "Sound spatialization as a navigational aid in virtual environments," in *6th CSI Crime Science and Investigation Conference*, Orlando, FL - USA, June, 2002.

[197] Dirk-Jan Kroon. (2011, January) Kinect Matlab. [Online]. http://www.mathworks.com/matlabcentral/fileexchange/30242-kinect-matlab

[198] M Everingham, B Thomas, and T Troscianko, "Head mounted mobility aid for low vision using scene classification techniques," *International Journal of Virtual Reality*, vol. 3, no. 1, pp. 3-12, 1999.

[199] O Javed, S Ali, and M Shah, "Online detection and classification of moving objects using progressively improving detectors," in *Conference in Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA - USA, June, 2005.

[200] W Shiuh-Ku , K Chung-Ming , and T Shu-Kang, "Video object tracking using adaptive Kalman filter," *Journal of Visual Communication and Image Representation*, vol. 17, no. 6, pp. 1190–1208, 2006.

[201] Z Kalal, J Matas, and K Mikolajczyk, "Online learning of robust object detectors during unstable tracking," in *On-line Learning for Computer Vision Workshop*, Kyoto - Japan, May, 2009.

[202] R E Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[203] Z Guangcheng, H Xiangsheng, Z Stan, and Y Li, "Boosting Local Binary Pattern (LBP)-Based Face Recognition," *Advances in Biometric Person Authentication*, vol. 3338, no. 1, pp. 179-186, 2005.

[204] L Rainer and M Jochen , "n Extended Set of Haar-Like Features for Rapid Object Detection," *IEEE ICIP 2002*, vol. 3, no. 1, pp. 900-903, 2002.

[205] L Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[206] P Viola and M Jones, "Rapid object detection using a boosted cascade of simple features," in *Conference in Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA - USA, June, 2001.

[207] L Bourdev, S Majil, and J Malik, "Describing people: A poselet-based approach to attribute classification," in *International Conference on Computer Vision*, Barcelona - Spain, October, 2010.

[208] L Eikvil, "OCR - Optical Character Recognition," Oslo, 1993.

[209] Google Drive. (2012, January) About Optical Character Recognition in Google Drive. [Online]. http://support.google.com/docs/bin/answer.py?hl=en&answer=176692

[210] R Minetto and N Thome, "Text Detection and Recognition in Urban Scenes," in *IEEE International Conference on Computer Vision Workshops*, New York - USA, May, 2011.

[211] A Cambra and A Murillo, "Towards robust and efficient text sign reading from a mobile phone," in *IEEE International Conference on Computer Vision Workshops*, New York - USA, May, 2011.

[212] P Yi-Feng , H Xinwen, and L Cheng-Lin, "Text Localization in Natural Scene Images Based on Conditional Random Field," in *10th International Conference on Document Analysis and Recognition*, Washington - USA, November, 2009.

[213] M Petter, V Fragoso, T Matthew, and B Charles, "Automatic text detection for mobile augmented reality translation," in *Intenational Conference in Computer Vision (ICCV) Workshops*, Barcelona - Sapin, October, 2011.

[214] Y Zhong, K Kalle, and J Anil, "Locating text in complex color images," *Pattern Recognition*, vol. 28, no. 10, pp. 1523–1535, 1995.

[215] G Lixu, T Naoki, R Haralick, and K Toyohisa, "The Extraction of Characters from Scene Image Using Mat hemat ical morphology," in *IAPR Workshop on Machine Vision Applicatio*, Tokyo - Japan, March, 1996.

[216] K Kim, K Jung, and J Kim, "Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1631-1639 , 2003.

[217] S Hanif and L Prevost, "Texture based text detection in natural scene images: a help to blind and visually impaired persons," in *Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments*, Glasgow - UK, September, 2007.

[218] M Anand, A Karteek, and C Jawahar, "Scene Text Recognition using Higher Order Language Priors," in *British Machine Vision Conference*, London - UK, September, 2012.

[219] L Neumann and J Matas, "Real-Time Scene Text Localization and Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RD - USA, June, 2012.

[220] L Jung-Jin, P Lee, S Lee, A Yuille, and C Koch, "Adaboost for text detection in natural scene," in *International Conference on Document Analysis and Recognition (ICDAR)*, Beijing - China, May, 2011.

[221] K Wang, B Babenko, and S Belongie, "End-to-end Scene Text Recognition," in *International Conference in Computer Vision (ICCV)*, Barcelona - Spain, October, 2011.

[222] P Felzenszwalb and D Huttenlocher, "Pictorial structures for object recognition," *IJCV*, vol. 61, no. 1, pp. 55–79, 2005.

[223] H Larochelle, Y Bengio, J Louradour, and P Lamblin, "Exploring Strategies for Training Deep Neural Networks," *Journal of Machine Learning Research*, vol. 20, no. 1, pp. 1-40, 2009.

[224] A Krizhevsky and G Hinton, "Using Very Deep Autoencoders for Content-Based Image Retrieval," in *European Symposium on Artificial Neural Networks ESANN*, Bruges - Belgium, October, 2011.

[225] R Sarikaya and G Hinton, "Deep Belief Nets for Natural Language Call-Routing," in *International Conference on Acoustics, Speech, and Signal Processing ICASSP*, New York - USA, November, 2011.

[226] P Baldi and K Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Networks*, vol. 2, no. 10, pp. 53–58, 1989.

[227] Y LeCun, L Bottou, Y Bengio, and P Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[228] Stanford University. (2011, April) UFLDL Tutorial. [Online]. http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

[229] G Bologna, B Deville, and T Pun, "On the use of the auditory pathway to represent image scenes in real-time," *Neurocomputing*, vol. 72, no. 1, pp. 839–849, 2009.

[230] G Barrow and M Tenenbaum, "Interpreting line drawings as three-dimensional surfaces," *Artificial Intelligence*, vol. 17, no. 3, pp. 75-116, 1983.

[231] J Ward and M Graham, "Evolving the ideal visual-to-auditory sensory substitution device using interactive genetic algorithms," *Journal of human, animal, and machine perception*, vol. 2, no. 8, pp. 37–41, 2011.

[232] M Auvray, S Hanneton, and K O'Regan, "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, 2007.

[233] J Ward and T Wright, "The evolution of a visual-to-auditory sensory substitution device using interactive genetic algorithms," *The Quarterly Journal of Experimental Psychology*, vol. 1, no. 4, pp. 37–41, 2013.

[234] G Bologna, B Deville, and T Pun, "Blind navigation along a sinuous path by means of the See ColOr interface," in *The 3rd international Work-Conference on the interplay between Natural and Artificial Computation*, Santiago de Compostela - Spain, June, 2009.

[235] V Weiss, S Cloix, G Bologna, and T Pun, "A robust, real-time ground change detector for a "smart" walker," in *the 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.*, Lisbon - Portugal, January, 2014.