This is an author manuscript post-peer-reviewing (accepted version) of the original publication. The layout of the published version may differ .

- -

On the use of the auditory pathway to represent image scenes in real-time

- -

Bologna, Guido; Deville, Benoît; Pun, Thierry

On the Use of the Auditory Pathway to Represent Image Scenes in Real-Time

G. Bologna[1], B. Deville[2], T. Pun[2]

[1]University of Applied Science
Rue de la prairie 4, 1202 Geneva, Switzerland
e-mail : guido.bologna@hesge.ch

[2]Computer Science Department, University of Geneva
Route de Drize 7, 1227 Carouge, Switzerland
e-mail : {benoit.deville, thierry.pun}@cui.unige.ch

Abstract

The See Color interface transforms a small portion of a coloured video image into sound sources represented by spatialised musical instruments. Basically, the conversion of colours into sounds is achieved by quantisation of the HSL colour system. Our purpose is to provide visually impaired individuals with a capability of perception of the environment in real time. In this work we present the system principles of design and several experiments that have been carried out by several blindfolded persons with See ColOr prototypes related to static pictures on a tablet and simple video images. The goal of the first experiment was to identify the colours of static pictures' main features and then to interpret the image scenes. Although learning all instrument sounds in only a training session was too difficult, participants found that colours were helpful to limit the possible image interpretations. The experiments on the analysis of static pictures suggested that the order of magnitude of the slow down factor related to the use of the auditory channel, instead of the visual channel could correspond to the order of magnitude related to the ratio of visual channel capacity to auditory channel capacity. Afterwards, two experiments based on a head mounted camera have been performed. The first experiment pertaining to object manipulation is based on the pairing of coloured socks, while the second experiment is related to outdoor navigation with the goal of following a coloured serpentine painted on the ground. The "socks" experiment demonstrated that blindfolded individuals were able to accurately match pairs of coloured socks. The same participants with the addition of a blind individual successfully followed a red serpentine painted on the ground for more than 80 meters. According to task time durations, the order of magnitude of the slow down factor related to the "socks" and "serpentine" experiments 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.

1   Introduction

Although the world we observe seems very stable, vision is a dynamic and fugitive phenomenon. Specifically, eye movements indicate the location of meaningful content in an image. The central area of the human retina (fovea) spans for approximately two degrees. To understand a scene image our eyes scan it in a series of fixations, without conscious planning. The rapid eye movements between fixations are designated as saccades. Note that the time duration of a fixation is related to the processing expended on interpreting the corresponding portion of the image [14]. Within each fixation, the fine detail that will be visible depends on its contrast, its spatial frequency and its eccentricity, or angular distance from the center of the field of view [21, 18].

Following a 2002 survey, the World Health Organisation estimated there were 161 million visually impaired people in the world, of whom 124 million had low vision and 37 million were blind [27]. For a blind person, the quality of life is appreciably improved with the use of special devices, which facilitate precise tasks of everyday life, such as reading, manipulating objects or using a computer. The most common modality to replace sight is touch, more precisely the haptic modality composed of two complementary channels, tactile and kinesthetic [20].

The cutaneous stimulations generated by tactile displays are based on three distinct classes of stimuli: mechanical tactile; electro-tactile and thermo-tactile [9]. Mechanical tactile substitution remains however the more widely used, thanks to its best spatio-temporal resolution. In this case, the stimuli are obtained by means of a vibro-tactile matrix, which is composed of a 2D lattice of tractors. The height of the tractors can vary according to several levels, in order to represent three-dimensional surfaces. Specifically, four criteria determine the quality of tactile displays: the number of tractors included in the device; the density of the tractors; the number of

vertical levels that the tractors can take and the refresh frequency. These devices, transform the visual flow collected by a mobile camera into tactile stimuli. Tactile displays are fixed on areas of the body having good spatio-temporal resolution like the abdomen [2], the face or more recently, the tongue [3].

Many experiments were carried out by Kaczmarek et al. on congenital blind men [15]. The results obtained with a vibro-tactile device showed that after a sufficiently long training period, the participants were able to provide certain perceptual impressions like depth and perspective. However, the results also revealed several deficiencies related to the size of the tactile display and its weak resolution.

Optacon II is a device including a photo-receiver connected to a vibro-tactile matrix made up of 144 tractors arranged in 24 lines and 6 columns. Its use implied to place the finger of a hand on the tactile display and to use the other hand to sweep the contents of the image or the text [24]. Experiments reported that blind subjects were able to recognize elementary patterns.

Kawai and Tomita developed a prototype for the recognition of voluminous objects [17]. The system, such as it was carried out brings into play two cameras, a touch screen and a sound helmet. The two cameras were used to obtain a stereographic image of the objects of the environment. The touch screen is composed of 16x16 tractors laid out on a surface of 175x175 mm. The space ranging between each tractor is 10 mm, while their height can vary between 0 and 6 mm. The synthetic voice is used to add additional information. Evaluations carried out on five visually impaired subjects showed that they were able to recognize without difficulties the objects of the everyday life (cup, ball, book, etc. …).

Maucher, developed a tactile device which includes 48 piezoelectric actuators laid out on a 4x12 matrix [22]. This matrix which functions as a pantograph is fixed by a rail on a shelf whose surface of work is of 220x220 mm. The refresh frequency of the tactile device is 20 Hz. The system was assessed by six blind subjects and four normal sighted persons on simple geometrical objects like squares, triangles or circles with varying size and orientation. The results showed that blind subjects recognised the objects twice faster than subjects in the normal control group.

The second most common modality to replace sight is audition. Echolocation is a mode of perception used spontaneously by many blind people. It consists in perceiving the environment by generating sounds and then listening to the corresponding echoes. Reverberations of various types of sound, such as slapping of the fingers, murmured words, whistles, noise of the steps, or sounds from a cane are commonly used.

Sound spatialisation is the principle which consists of virtually creating a three-dimensional auditory environment, where sound sources can be positioned all around the listener. These environments can be simulated by means of loudspeakers or headphones. Among the precursors in the field, Ruff and Perret led a series of experiments on the space perception of auditory patterns [25]. Patterns were transmitted through a 10x10 matrix of loudspeakers separated by 10 cm and located at a distance of 30 cm from the listener. Patterns were represented on the auditory display by sinusoidal waves on the corresponding loudspeakers. The experiments showed that 42% of the participants identified six simple geometrical patterns correctly (segment of lines, squares, etc). However, orientation was much more difficult to determine precisely. Other experiments carried out later by Lakatos taught that subjects were able to recognise with 60-90% accuracy ten alphanumeric characters [19].

Several authors proposed special devices for visual substitution by the auditory pathway in the context of real time reactivity. The "K Sonar-Cane" combines a cane and a torch with ultrasounds [16]. Note that with this special cane, it is possible to perceive the environment by listening to a sound coding the distance.

"TheVoice" is another experimental vision substitution system that uses auditory feedback. An image is represented by 64 columns of 64 pixels [23]. Every image is processed from left to right and each column is listened to for about 15 ms. In particular, every pixel gray level in a column is represented by a sinusoidal wave with a distinct frequency. High frequencies are at the top of the column and low frequencies are at the bottom in a 5 kHz range of frequencies.

Capelle et al. proposed the implementation of a crude model of the primary visual system [11]. The implemented device provides two resolution levels corresponding to an artificial central retina and an artificial peripheral retina, as in the real visual system. The auditory representation of an image is similar to that used in "TheVoice" with distinct sinusoidal waves for each pixel in a column and each column being presented sequentially to the listener. Experiments carried out with 24 blindfolded sighted subjects revealed that after a period of time not exceeding one hour, subjects were able to identify simple patterns such as horizontal lines, squares and letters.

A more musical model was introduced by Cronly-Dillon [12]. First, the complexity of an image is reduced by applying several algorithms (segmentation, edge detection, ...). After processing, the image contains only black pixels. Pixels in a column define a chord, while horizontal lines are played sequentially, as a melody. When a processed image presents too complex objects, the system can apply segmentation algorithms to these complex objects and to obtain basic patterns such as squares, circles and polygons. Experiments carried out with normal

and (elderly) blind persons showed that in many cases a satisfactory mental image was obtained.

Gonzalez-Mora et al. developed a prototype using the spatialisation of sound in the three dimensional space [13]. The sound is perceived as coming from somewhere in front of the user by means of head related impulse responses (HRIRs). The first device they achieved was capable of producing a virtual acoustic space of 17x9x8 gray level pixels covering a distance of up to 4.5 meters.

In this paper, we present *See ColOr* (Seeing Colors with an Orchestra), which is an ongoing project aiming at providing visually impaired individuals with a non-invasive mobility aid. See ColOr uses the auditory pathway to represent in real-time frontal image scenes. General targeted applications are the search for items of particular interest for blind users, the manipulation of objects and the navigation in an unknown environment. The See ColOr interface encodes coloured pixels by spatialised musical instrument sounds, in order to represent and emphasize the color and location of visual entities in their environment [6, 7, 8]. The basic idea is to represent a pixel as a directional sound source with depth estimated by stereo-vision. Finally, each emitted sound is assigned to a musical instrument, depending on the colour of the pixel.

In this article are reported the experiments that have been carried out by several blindfolded persons with See ColOr prototypes related to static pictures on a tactile tablet and simple video images. The goal of the first experiment was to identify the colours of the main features of static pictures and then to give an interpretation of the whole image scenes. Experiment participants found that colours were helpful to limit the possible image interpretations. The experiments on the interpretation of static pictures suggested that the order of magnitude of the slow down factor related to the use of the auditory channel, instead of the visual channel could correspond to the order of magnitude related to the ratio of visual channel capacity to auditory channel capacity, which represents two orders of magnitude.

In addition, two experiments based on head mounted cameras have been performed (a video is available on http://129.194.70.56/see_color_demos). We aimed at verifying the hypothesis that it is possible to manipulate and to match coloured objects, such as coloured socks with an auditory feed-back represented by sounds of musical instruments. The last purpose was to verify that navigation in an outdoor environment can be performed with the help of the sound related to a coloured serpentine painted on the ground. The "socks" experiment demonstrated that blindfolded individuals were able to accurately match pairs of socks. The same participants and a blind person successfully followed a red serpentine for more than 80 meters. To the best of our knowledge this is the first study in the context of visual substitution for object manipulation and real time navigation, for which colour is supplied to the user as musical instrument sounds. In the following sections we present in section 2 the sound code, section 3 describes the three See ColOr prototypes, section 4 examines several considerations about the coded amount of information, and section 5 illustrates the experiments, followed by the conclusion.

2   Sound Code

Basically, the conversion of sounds into colours is achieved by quantisation of the HSL colour system. Subsequently, the obtained sounds are spatialised by means of HRIRs.

2.1   Colour system conversion

The goal is to use the auditory pathway to convey colour information as quickly as possible. The simplest method would consist to use human voice. The main problem is that we would like to communicate several pixel colours, simultaneously. Note that for a person it is almost impossible to follow at the same time a discussion with more than three individuals. It may be possible to understand the name of two colours, though saying a colour takes about a second or even more if the palette has a considerable size, which is also too long for real-time purposes.

Another approach consists in the sonification of colour system variables. The RGB (red, green, and blue) cube is an additive colour model defined by mixing red, green and blue channels. One important drawback of this model is that similar colours at the human perceptual level could result considerably further on the RGB cube, which would generate perceptually distant sounds.

As shown by Figure 1, the HSL colour system (Hue, Saturation and 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 colours. The *H* variable represents hue from red to purple (see Figure 2), the second one is saturation which represents the purity of the related colour and the third variable represents luminosity. Hue varies between 0 and 360 degrees, 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 colour of music lives in the timbre of performing instruments. Moreover, learning to associate instrument timbres to colours 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.
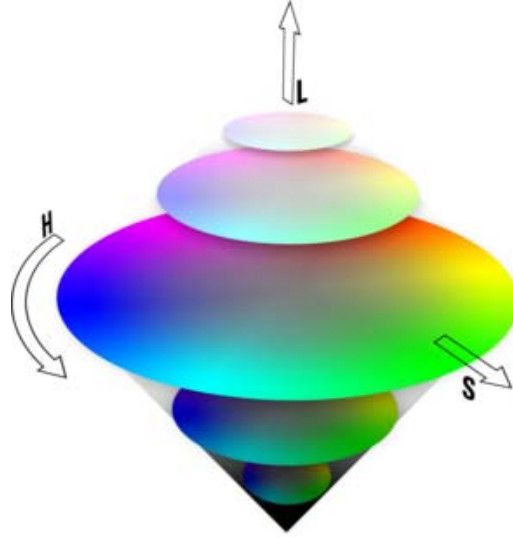
Figure 1. The HSL colour system; the distinct colour variables are hue (H), saturation (S) and luminosity (L).
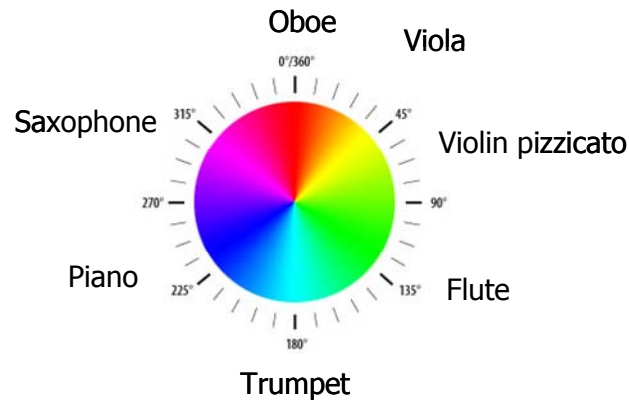
Figure 2. The association between hue and classical instrument sounds.

With respect to the hue variable, our empirical choice of musical instruments is:

- 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

4

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

with g representing the gain defined by

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

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 for ($0 \leq S < 0.25$);
- Sol for ($0.25 \leq S < 0.5$);
- Si flat for ($0.5 \leq S < 0.75$);
- Mi 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 for $0 \leq L < 0.125$
- Sol for $0.125 \leq L < 0.25$
- Sib for $0.25 \leq L < 0.375$
- Mi 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 for $0.5 \leq L < 0.625$
- Sol for $0.625 \leq L < 0.75$
- Sib for $0.75 \leq L < 0.875$
- Mi for $0.875 \leq L \leq 1$

Moreover, if luminance is close to zero, the perceived colour 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 colour 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.

2.2 Sound spatialisation

It is possible to simulate lateralisation, also denoted as two-dimensional auditory spatialisation, 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 spatialisation. In fact, the folds of the pinnae cause echoes with minute time delays within a range of 0-0.3 ms [5] 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 HRIRs, which not only varies in a complex way with azimuth, elevation, range, and frequency, but also varies significantly from person to person [10].

Generally, reproducing lateralisation with uncustomised 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 lateralisation on a row of 25 points with the use of the CIPIC database [1]. In practice, each sonified point corresponds to the convolution of an instrument sound with the

corresponding HRIR filter related to a particular azimuth position.

3   See ColOr Prototypes

We implemented three distinct prototypes. As shown by Figure 3, the first is related to a tactile tablet with the sonification of static pictures, while the others sonify pixel points captured by a stereoscopic camera (Figure 4) and also by a webcam (Figure 5).

Figure 3. The See ColOr tactile prototype

Figure 4. The See ColOr stereoscopic camera prototype.

Figure 5. The See ColOr webcam prototype.

3.1   Tactile tablet prototype

The tactile prototype is based on the T3 tactile tablet from the Royal National College for the Blind (http://www.talktab.org/). We put on the T3 tablet a special paper with images including detected edges represented by palpable roughness. This device allows to point on a picture with the finger and to obtain the coordinates of the contact point. A row of 25 points with the middle point represented by the coordinates of the touched point is sonified for 300 ms by means of spatialised instrument sounds. In order to replicate a crude model of the visual system, we have maximal resolution at the centre of the row and lower resolution in the periphery. In practice we skip a given number of pixels. As shown below, starting from the middle point (in bold), the following vector of 25 points represents the number of omitted pixels.

$$[14\ 11\ 8\ 6\ 4\ 2\ 2\ 1\ 1\ 0\ 0\ 0\ \mathbf{0}\ 0\ 0\ 0\ 1\ 1\ 2\ 2\ 4\ 6\ 8\ 11\ 14]$$

The HRIR filters we use for spatialisation are those included in the CIPIC database. Specifically, measurements of the KEMAR manikin [1] are those used by our See ColOr interface. All possible spatialised sounds (25*9*4 = 900) are pre-calculated and reside in memory. In practice, our main program for sonification is a mixer selecting appropriate sounds, with respect to the contact point of the finger on the picture. The maximal time latency for generating a sonified row of 25 points is 50 ms with the use of Matlab on a Pentium 4 at 3.2 GHz.

3.2   Stereoscopic camera prototype

We use a stereoscopic colour camera denoted STH-MDCS2 (SRI International: http://www.videredesign.com/). An algorithm for depth calculation based on epipolar geometry is embedded within the stereoscopic camera. The resolution of images is 320x240 pixels with a maximum frame rate of 30 images per second. As with the tactile tablet, we sonify a row of 25 pixels with spatial lateralisation by means of the CIPIC database and also high resolution in the middle and low resolution in the periphery. The time latency is 80 ms, which is 30 ms more than with the previous prototype, because of the image capture. In practice, with a sonification length of 300 ms and with a time latency of 80 ms the sonification frequency is about 2.5 Hz.

Recently, we have implemented the sonification of depth; because of the limited bandwidth of the auditory channel (cf. section 4) we only take into account the sonification of a particular area of this row. Specifically, we first determine among these 25 points the maximal number of contiguous points labelled with the same colour. Then, we calculate the centroid of this area and the average depth. It is possible to have points of undetermined depth, especially in homogeneous areas like walls, for which the depth algorithm is unable to determine remarkable points related to the calculation of the disparity between the left and right images. Points with undetermined depth are not considered in the average depth calculation. The final sonification presents only a spatialised sound source representing the average colour and the average depth. In the future, it will be possible

to switch from the depth mode to the normal mode (not including depth encoding), with the use of a mouse button.

At the auditory level the depth variable is represented by sound time duration between 0 and four meters, and by volume lessening from four meters to infinity. The correspondences between sound duration and average sound depth D are :

- 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 ($3 \leq D \leq 4$).

Then, after four meters the volume V starts to decrease by following a negative exponential function given by

f(V) = V*exp(-k*D);

with k a positive small constant.

3.3   Webcam prototype

This prototype is based on a Webcam which can adapt efficiently to light changing conditions related to outdoor purposes. We use the Logitech Notebook Pro webcam with the RightVision2 technology. We implemented the sonification of a sub-window containing 17x9 points by means of streaming. Nevertheless, due to auditory channel limitations, usually we configure the sonification exclusively for a row of 17 pixels with constant resolution and by sonifying a pixel out of four, with respect to the centre of the video image. Note that here depth is not taken into account. The spatialisation is achieved by virtual ambisonic of order two [6] with CIPIC Kemar HRIRs. Note that for sonification configurations with more than a sonified row, the sound is spatialised by means of elevation. The time latency of the C++ implemented program is about a few milliseconds. Finally, the sonification speed can be chosen between three and eleven Hz.

4   Mobility Aid Channel Capacity

Our senses can be regarded as data channels possessing a maximal capacity measured in bits per seconds (bps). The bandwidth of the human auditory system is about 20 Khz, thus by virtue of the Nyquist theorem the maximal amount of information that could be transmitted through this channel without loss is 40 Kbps. Note that the visual channel has the largest capacity with about 1000 Kbps [26], while the sense of touch would is less powerful than audition with about 0.1 Kbps [26]. Total vision substitution would be impossible to achieve, as the bandwidth of touch or audition would be inadequate. Therefore, it is important to make crucial choices in order to convey a small part of vision, in such a way that the reduced visual information rendered by another sense preserves useful meaning.

Let us estimate the amount of information transmitted by several mobility aids. "TheVoice", with the sonification of 64x64 pixels quantised on 16 grey levels and sonified for about a second corresponds to approximately 65 Kbps. Gonzalez-Mora et al. prototype represents16x9x8 points with grey levels coded by the volume gain. Assuming the use of 8 quantised volume levels, the required capacity would be about 9 kbps. Moreover, we must take into account that the output sounds correspond to small clicks of 0.5 ms repeated 25 times per second (sound example : http://eav.ull.es/sonido.html). Thus, the amount of information transmitted is 225 Kbps.

Our see ColOr interface takes into account colour. We use two instrument timbres for the hue variable H; although we use a continuous coding of H, assuming that between two instruments we are able to distinguish four distinct timbre combinations, we obtain 7*4=28 possibilities. The saturation variable S depends on four quantised levels and the luminance L takes 8 different values. Thus, we have 28*4*8 = 896 different colours. We must also consider the addition of eight grey levels that arise in situations of low/high luminance for which every colour becomes grey/black/white. Thus, we obtain 896+8=904 possible colours for a single pixel. Sonifying a row of 25 points would involve a required capacity of about 56.5 Kbps at 2.5 Hz sonification rate (cf. stereoscopic prototype). The depth mode is only applied to a single sound source, thus with five distinct sound

time durations and five distinct volume levels, the transmitted information is about 9 Kbps.

The amount of transmitted information of the mobility aids described above is beyond their corresponding maximal capacity thresholds. In fact, the maximal capacity of "TheVoice" is 10 Kbps, as its sound bandwidth is 5 KHz. Assuming a 20 KHz bandwidth, Gonzalez-Mora et al. prototype is also beyond the maximal capacity threshold, as the amount of transmitted information is 225 Kbps.

For our See ColOr prototypes, the use of musical instrument sounds covering the human auditory spectrum with their corresponding fundamental frequencies and harmonics allows us to take advantage of the 20 KHz bandwidth of the human auditory system. The See ColOr transmitted information exceeds the auditory channel capacity by a factor of about 1.4. However, it is important to consider that in many realistic situations close pixels are correlated. For instance, when looking at a landscape, several homogeneous regions are present and very often the sonified pixels would represent a few different colours, thus requiring less channel capacity. Note that situations involving channel capacity overload would be analogous to coloured pepper and salt noise, which are rare in typical image scenes. Nevertheless, a remedy would correspond to the listening to the same sonified region several times and focusing the attention on smaller sonified portions, such that the missing bandwidth would be compensated by sonification repetitions.

We can also reduce the quantity of transmitted information by sonifying a smaller amount of pixels, as in our webcam prototype comprising a sonified row of 17 points, conveying 38.4 Kbps at a sonification rate of 2.5 Hz. However, it is important to make available a crude model of peripheral vision, thus we prefer the sonification configuration related to 25 pixels, as with more sonified points we better emulate the mechanism of peripheral vision.

5   Experiments

This section presents the experiments related to static pictures and video images captured by cameras. For each experiment we calculated the statistic based on simple population averages.

5.1   Static sonified pictures

The purpose of this study is to investigate whether individuals can learn associations between colours and musical instrument sounds. Another issue is whether it is possible to interpret pictures. Several experiments were carried out by six participants having their eyes enclosed by a dark tissue, and listening to the sounds via headphones [7, 8]. In these experiments we use the T3 tactile tablet (cf. section 3.1).

Six participants were trained to associate colours with musical instruments and then asked to determine on several pictures, objects with specific shapes and colours. Experiments involved a training phase with the use of elementary pictures. For all our experiment participants, training lasted about 45 minutes. Afterwards, a small test for scoring the performance of the participants on sound/colours associations was achieved. On the 15 heard sounds, the average number of correct colours among the six participants was 8.1 (standard deviation : 3.4). It is worth noting that the best score was reached by a musician who found 13 correct answers. Afterwards, participants were asked to explore and identify the major components of several pictures.

Regarding the children draw picture illustrated in Figure 6, all participants interpreted the major colours 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.
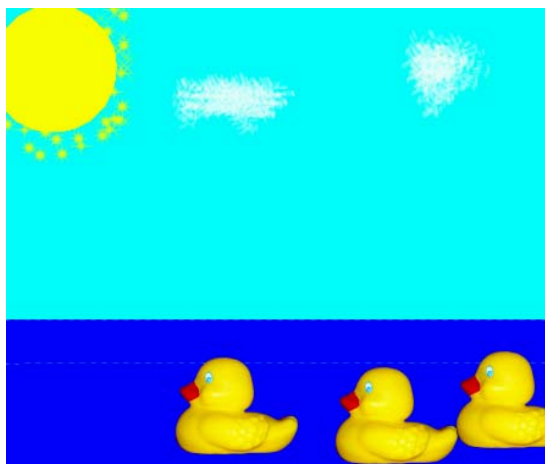
Figure 6. The "ducks" picture; the goal was to determine the colour of the main components (sun, sky, sea, duks, etc.). and to interpret the picture.

For the picture depicted in Figure 7 all participants interpreted the major colours as the sky and the sea; an individual said that the dolphin is 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.
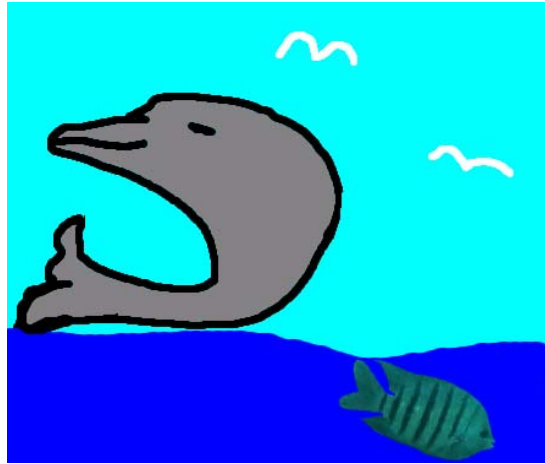
Figure 7. The "dolphin" picture; the goal was to determine the colour of the main features (sky, sea, dolphin, etc.) and to interpret the picture.

On the interpretation of real images, such as the picture shown in Figure 8, 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 !


Figure 8. The "tree" picture. The goal was to interpret this real world image scene.

Concerning Figure 9, all subjects found major colours (blue and yellow); however no one made the distinction between the sky and the sea. Note however that two participants suggested that the grey/white area between these two components represented clouds. Interestingly, three participants correctly analysed the luminosity of the sand, which corresponds to a low saturated yellow and two of them determined that the sky was more saturated at the top with progressive decreasing saturation levels and increasing luminosity when moving to the grey/white area. The yellow cliff was not identified, but an individual proposed a hay stack. In fact pictures with a perspective view are too difficult to be interpreted.

Figure 9. The "cliff" picture. The goal was to interpret the image scene.

The last assignment was to find a red door in Figure 10. The two red doors represent less than 1% of the picture surface. As shown by Table 1, which summarizes the time durations for the exploration of pictures, all participants found one of the red doors in a time range between 4 and 9 minutes. The average time being 6.6 minutes and given the hypothesis that normal sighted individuals would determine one of the red doors in at most a second, the slow down factor with the use of See ColOr would correspond to two orders of magnitude, which is consistent to the ratio of visual channel capacity to auditory channel capacity. Moreover, this ratio roughly appears during the identification of the main components of the other pictures. In fact, the average time of all the participants to explore Figures 6, 7, 8, and 9 is 9.2 minutes, while in at most two seconds an individual would identify relevant image features.
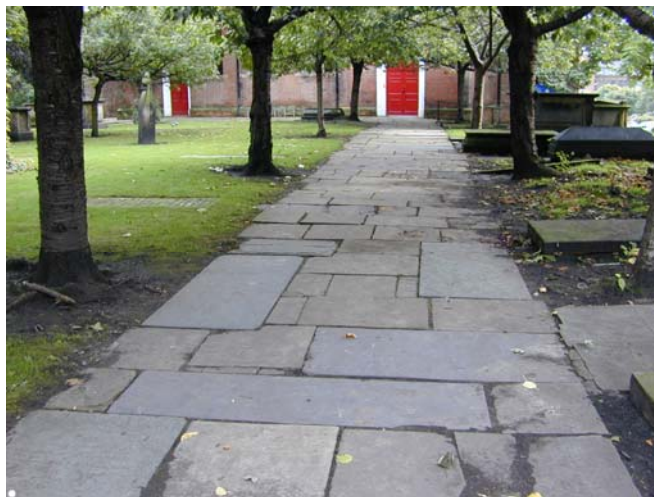
Figure 10. The "churchyard" picture; the goal was to find a red door.

Participant	Fig. 6	Fig. 7	Fig. 8	Fig. 9	Fig. 10
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 1. Time durations in minutes for the examination of several pictures.

5.2 Pairing coloured socks

The purpose is to verify the hypothesis that with the use of a camera, it is possible to manipulate and to match coloured objects with an auditory feed-back represented by sounds of musical instruments. As shown in previous section, it is difficult to learn the associations between colours and sounds in just one training session, thus our participants were not asked to identify colours, but just to pair similarly coloured socks. The camera here is the STH-MDCS2. Although we use a stereoscopic camera, depth is not sonified.

The experiments are performed by seven blindfolded adults who were not present in the previous experiments, except one of them (P_3). The training phase includes two main steps. First, we explain associations between colours and sounds in front of a laptop screen showing different static pictures. Specifically, we show the HSL system with seven main hues and several saturation varying pictures. We let our participants decide when they feel comfortable to switch to the second step aiming at learning to point the camera toward socks. With respect to each individual, Table 2 illustrates the time dedicated for the two training steps.

Participant	Static Training (mn)	Training with Socks (mn)
P1'	12	12
P2'	7	11
P3	18	15
P4'	0	6
P5'	0	24
P6'	29	18
P7'	16	16
Average	11.7 ± 10.4	14.6 ± 5.7

Table 2. Training time durations without camera and with a head mounted camera pointing at socks.

As shown by Figure 11 we use five pairs of socks having the following colours: black, green, low saturated yellow, blue and orange. Table 3 illustrates the results of our experiment participants and Figure 12 shows an individual examining a sock. It is worth noting that the average number of paired socks is high. Participant P_4 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 colours 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.

Figure 11. The colored socks; from left to right : black, green, yellow, blue and orange.

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	**11.0 ± 5.5**	**4.7 ± 0.8**

Table 3. Testing time duration and success rate of the "socks" experiment with the See ColOr interface.

Figure 12. A blindfolded individual observing an orange sock with a head mounted camera.

A normal sighted person can match five pairs of coloured socks picked up from a plastic bag in 25 seconds, on average. Consequently, the slow-down average factor with the use of the See ColOr prototype is 26.4 times. Accordingly, this corresponds to roughly an order of magnitude. Note that even if we consider the best time of four minutes performed by participant P_2, we would obtain the same order of magnitude.

A question arising is the influence of training on the time required to pair socks. In fact, one of the authors

13

who is very well trained can perform this task in 2.2 minutes, which is almost twice faster than the best participant time duration. This represents also time duration of about 5 times slower than that obtained by normally sighted individuals.

5.3 Pairing coloured socks without See ColOr

Here the goal is to match coloured socks without the camera. In fact, we would like to demonstrate that just relying on touch involves random results. Table 4 illustrates the obtained results. With See ColOr the average matching accuracy was 94%, while without See ColOr the overall matching accuracy was 12%. Thus, the See Color interface significantly helped blindfolded participants to identify pairs of socks.

Participant	Time (mn)	Success Rate (pairs)
P1'	3	0
P2'	2	0
P3	13	1
P4'	4	1
P5'	2	0
P6'	4	2
P7'	7	0
Average	**5.0 ± 3.9**	**0.6 ± 0.8**

Table 4. Testing time duration and success rate of the "socks" experiment without the See ColOr interface.

The chance probability to pair the second sock correctly is 1/9 = 11.1%. After the second guess the calculation of probabilities is based on conditional probabilities. In other words, the probability to guess a pair of socks depends also on previous matching. For instance, if the second sock was not matched with the first, then the probability to have a matched sock for the fourth guess is (6/8)*(1/7) = 10.7%. Because of the numerous cases and the strong dependence of previously selected socks, it is clearly complex to give a precise probability of the five pairs' matching, however it is possible to estimate an upper bound.

For the first pair this probability is exactly 1/9. For the second pair the upper bound probability is 1/7. Similarly, for the third and fourth pairs the probability is bounded by 1/5 and 1/3, respectively. Finally, for the last pair the probability of correct matching is 1/9. More precisely, there are 45 combinations of eight socks out of ten. In these 45 combinations, we find five combinations leaving out two identical socks (5/45 = 1/9). Therefore, the overall probability of matching a pair of socks during an experiment is bounded by 1/9 + 1/7 + 1/5 + 1/3 + 1/9 = 89.8%. In other words, a thousand individuals would match at most 898 pairs of socks. For the experiment without the See ColOr interface we counted four matched pairs. Since the number of participants is equal to seven, we should obtain at most six matched pairs. As we counted four matched pairs (see table 3), the results can be regarded as random.

5.4 Following a coloured serpentine

The purpose is to verify the hypothesis that it is possible to use the See ColOr interface to follow a coloured line or serpentine painted on the ground of an outdoor environment. Figure 13 illustrates an individual performing this task. For this experiment we incorporate the same seven individuals who carried out the experiment on coloured socks and a blind person. The camera here is the Logitech Quickcam Notebook Pro.

The training phase lasts approximately ten minutes. Specifically, a supervisor manages an experiment participant in front of the coloured serpentine. The experimenter is 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). The sonification frequency is fixed to 4 Hz. For experienced users it would be possible to increase the frequency at the maximal implemented value of 11.1 Hz. Afterwards, we ask to the person performing the experiment to move the head from left to right and to become aware that the oboe sound shifts. Note that the supervisor wears a headphone and can listen to the sounds of the interface. Finally, the experimenter is asked to start to walk and to keep the oboe sound in the middle sonified region. Note that the training session is quite short. An individual has to learn to coordinate three components. The first is the oboe sound position (if any), the second is related to the

awareness of the head orientation and the third is the alignment between the body and the head. Ideally, the head and the body should be aligned with the oboe sound in the middle.

Figure 13. A blindfolded individual following a coloured line with a head mounted webcam and a notebook carried in a shoulder pack.

The purpose of the test is to go from a starting point S to a destination point T. The testing path is different from the training path. Several small portions of the main path M can be walked through three possible alternatives denoted as A, B, and C. The shortest path M has length of more than 80 meters. It is important to note that it is impossible to go from S to T by just moving straight ahead. In Table 5 we give for each experiment participant the training time duration and the testing time duration, while Table 6 illustrates the followed length path and the average speed. All our experiment participants reached point T from point S and no-one was lost and asked to be helped. One of the authors who is very well trained to the ColOr sound code went from S to T through the path $M+C$ in 4.2 minutes, corresponding to a speed average of 1257 m/h. Therefore, "novice users" could potentially improve their average speed after several training sessions.

Participant	Training Time (mn)	Testing Time (mn)
P1'	11	7.3
P2'	10	7.1
P3	8	13.6
P4'	9	8.5
P5'	10	10.4
P6'	10	9.7
P7'	10	12.9
Average	**9.7 ± 0.9**	**9.9 ± 2.6**

Table 5. Training and testing time duration of blindfolded individuals following a red serpentine.

Participant	Path Length (m)	Speed Average (m/h)
P1'	M+C = 88	723
P2'	M = 84	710
P3	M+B = 110	485
P4'	M+A = 93	656
P5'	M = 84	484
P6'	M+A+C = 97	600
P7'	M+A+C = 97	451
Average	**93.3 ± 9.2**	**587.0 ± 114.1**

Table 6. Path length and speed average of blindfolded individuals following a red serpentine.

15

A blind person participated in this experiment. He successfully learned to follow the red serpentine in 5.3 minutes. During the testing phase he went from *S* to *T* along the main path in 6.1 minutes. Thus, his average speed was 826 m/h, which is better than that measured on our best inexperienced blindfolded participant.

Concerning the slow down factor related to the use of the auditory channel, we think that even if very difficult, a normal sighted sprinter would be able to follow this serpentine up to a speed of 20-30 Km/h. Thus, we suggest that the order of magnitude of the slow down factor is again equal to one.

5.5 Discussion

We used simple images like those representing children drawings, as well as simple video scenes showing coloured socks and red serpentines. The ultimate goal being mobility in unknown environments, we support the view that the user focus of attention could choose relevant information from the sound code of coloured 3D images. Specifically, when one tries to identify a picture component, the audio representation of colour and the position in the scene limit the number of possible interpretations. For instance, when one looks to the ground and it "sounds green" (by means of a flute sound), it is very likely that what one is looking at is grass.

We prefer to avoid the explicit simplification of video images, because image/video processing algorithms, in order to be efficient have to be fine-tuned for particular situations. In addition, with the current state-of-the-art, it is not clear at all that such algorithms could be used in real life mobility situations. Finally, video simplification requires significant processing time. As a consequence, we might expect poor results in terms of the quality of the interpretation by the user, for instance under different illumination conditions or for complicated scenes. Processing time would hamper reactivity in situations where it is required, such as navigation in unknown environments. Note also that the implicit simplification performed by the quantization of the HSL colour space is computationally efficient.

As a possible usefulness of the See ColOr system, we could imagine a blind individual following a path painted on the ground in an indoor environment, such that of a shopping center or of a medical center. In practice, in a complicated environment a blind person can get lost very easily; thus the painted line would be very helpful to overcome this problem. Similarly, this could be applied to a garden park or to a sidewalk leading to specific interest places. Moreover, for several points on a path it could be interesting to complement the auditory rendering by conveying specific information with RFID's or informative panels that could be read by a computer. Since the cost of the See ColOr prototype with a webcam and also the cost of a painted line on the ground are cheap, it would be economically advantageous for both blind individuals and public authorities to support such a framework.

A drawback of the See ColOr perceptual mode is the small angle of view that makes it difficult to perceive the local context. A remedy was to define the photographic mode, in order to enumerate the most important coloured regions. Another alternative could be the use of a small tactile device, such a touchpad. Since the amount of information that could be transmitted to the user is two orders of magnitude lower than that conveyed by the auditory channel, only specific visual features should be presented to the user. For instance, it could be possible to represent the obstacles from the user up to 10 meters. In practice, a blind individual would sense this tactile map with a hand and would receive a special sound when touching a point related to an obstacle.

It would have been problematic to follow the red serpentine solely by means of luminosity. For instance Gonzalez and Mora's prototype [13] and "TheVoice" [23] rely on that variable for their sound encoding. However, the luminosity of the red serpentine and the background are almost identical, therefore the user would not be able to perceive it with these prototypes.

As our ultimate goal is real time navigation, purely tactile rendition of scenes using dynamic displays suffers from several drawbacks. As previously discussed in section 4, the information transfer capacity of the tactile channel is inherently too limited. Moreover, such displays are technologically difficult to realize and costly; they also are difficult to use for extended periods of time, especially for those with vibrating pins. Finally, image/video scene simplification is needed, which is difficult to achieve with real scenes.

Invasive devices, such as retinal implants [28] raise a number of questions, such as long-term user acceptance, ethical concerns, technical feasibility, and cost. In addition to medical issues, these devices require that at least parts of the visual pathways are still operating: the optical nerve in case of artificial retina, as well as the visual cortex. We advocate here a non-invasive and low cost approach. Note also that these implants generate small visual percepts known as phosphenes, which appear as light spots, but without colour. It is worth noting that for the time being it is completely unknown how to elicit colour perception by means of electrode stimulation.

The reactivity of the See ColOr interface is important for tasks requiring real time constraints. An alternative to the colours coded by musical instrument sounds would be the use of colour names. However, the typical time duration would be in this case a second; hence, the reactivity of a system would be worse than that related to the

use of sounds lasting at most 300 ms, as in See ColOr. Furthermore, it would be difficult to remember the names of 904 different colours ! Generally, our senses detect subtle changes; in particular See ColOr allows the user to perceive the environment with its progressive variations that can be detected in the sound code by changing notes, modifications of instrument gains, progressive presence/absence of bass/singing voice in dark/bright regions, etc. In other words, with the use of instrument sounds a user can focus his/her attention to gradual changes, instead of listening to the true names of colours.

The experiments presented several tasks for which the interpretation of unknown image scenes was the most difficult. We conjectured in this case a slowdown factor of two orders of magnitude, while for head mounted camera experiments this variable would be equal to one. A question arising is whether it would be possible with a substantial number of training sessions to improve the performance by an order of magnitude. For the "serpentine" experiment it would imply that a blind individual could run on it. It seems to be unlikely to occur; however just walking at a normal speed would be great and sufficient for real life activities. Finally, another question arising is whether the interpretation of image scenes using a head mounted camera could be sped up by an order of magnitude with a user well acquainted with the See ColOr sounds of the environment.

6 Conclusion

We presented the current state of the See ColOr project, which provides the user with an auditory feedback of the colours of the environment. The purpose of the first experiment was to identify the colours of the main features of static pictures and then to understand image scenes. Although learning all instrument sounds in only a training session was too difficult, participants found that colours were helpful to limit the possible image interpretations. The experiments on the analysis of static pictures suggested that the order of magnitude of the slow down factor related to the use of the auditory channel, instead of the visual channel could 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.

With a camera prototype we verified the hypothesis that it is possible to manipulate and to match coloured objects, accurately. Overall, with only one training session, participants paired sock pairs with very high accuracy. With the last experiment related to blind navigation aiming at following a red serpentine painted on the ground, we validated the hypothesis that it is possible to follow a twisting path. The "socks" and "serpentine" experiments suggested that with the use of the See ColOr interface the order of magnitude of the slow down factor related to the use of the auditory channel, instead of normal vision would 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.

In the future we will pursue the socks' and serpentine experiments, in order to obtain more robust statistics. As well as blindfolded individuals, we are confident that blind persons will successfully follow the red serpentine, because of their improved auditory sense. Moreover, we will plan an experiment, for which depth represents an important parameter. In particular, we could imagine the presence of obstacles with an experimenter trying to estimate the distance separating him/her to an obstacle without touching it.

Aknowledgments

References

1. Algazi, V.R., Duda, R.O., Thompson, D.P., and Avendano, C. The CIPIC HRTF Database. In *Proc. WASPAA'01,* New Paltz, NY (2001).

2. Bach-y-Rita, P., Collins, C.C., Saunders, F.A., White, B., and Scadden, L. Vision substitution by tactile image projection. *Nature*, Vol. 221 (1969), 963–964.

3. Bach-Y-Rita, P. Visual information through the skin: a tactile vision substitution system (TVSS). Symp. on Prosthetic Aids for the Blind. *Trans. Amer. Acad. Ophtalm.,* Vol. 78, Sept. 1974.

4. Bach-y-Rita, P., Kaczmarek, K., Tyler, M., Garcia-Lara, J. Form perception with a 49-point electrotactile stimulus array on the tongue. *J. Rehab. Res. Develop., 35* (1998), 427–431.

5. Begault, R. 3-D Sound for Virtual Reality and Multimedia. Boston A.P. Professional (1994), ISBN: 0120847353.

6. Bologna, G., Vinckenbosch, M. Eye tracking in coloured image scenes represented by ambisonic fields of musical instrument sounds. In *Proc. International Work-Conference on the Interplay between Natural and Artificial Computation (IWINAC)*, (1) 327–333, Las Palmas, Spain, June 2005.

7. Bologna, G., Deville, B., Pun T., and Vinckenbosch, M. Identifying major components of pictures by audio encoding of colors. In Proc. *IWINAC'07* (2007), (2) 81–89.

8. Bologna, G., Deville, B., Pun T., and Vinckenbosch, M. Transforming 3D coloured pixels into musical instrument notes for vision substitution applications. *Eurasip J. of Image and Video Processing,* A. Caplier, T. Pun, D. Tzovaras, Guest Eds., (2007), Article ID 76204, 14 pages (Open access article).

9. Burdea, G.C. Force and touch feedback for virtual reality. John Wiley & Sons, Inc., New York, NY, 1996.

10. Brown, C.P., and Duda, R.O. A Structural Model for Binaural Sound Synthesis. *IEEE Trans. Speech and Audio Processing* (1998)*, 6* (5).

11. Capelle, C., Trullemans, C., Arno, P., and Veraart, C. A real time experimental prototype for enhancement of vision rehabilitation using auditory substitution. *IEEE T. Bio-Med Eng.,* 45 (1998), 1279–1293.

12. Cronly-Dillon, J., Persaud, K., and Gregory, R.P.F. The Perception of Visual Images Encoded in Musical Form: a Study in Cross-Modality Information. In *Proc. Biological Sciences* (1999) 266, 2427–2433.

13. Gonzalez-Mora, J.L., Rodriguez-Hernandez, A., Rodriguez-Ramos, L.F., Dfaz-Saco, L., and Sosa, N. Development of a new space perception system for blind people, based on the creation of a virtual acoustic space. In *Proc. IWANN'99* (1999), 321–330.

14. Just, M.A., and Carpenter, P.A. Eye fixations and cognitive processes. *Cognitive Psychology* (1976), 8, 441–480.

15. Kaczmarek, K., Bach-Rita, P., Tompkins, W.J., and Webster, J.G. A tactile vision-substitution system for the blind: computer-controlled partial image sequencing. *IEEE Trans. Biomed. Eng., 32* (1985), 602–608.

16. Kay, L. A sonar aid to enhance spatial perception of the blind: engineering design and evaluation. *The Radio and Electronic Engineer,* 44 (1974), 605–627.

17. Kawai, Y., and Tomita, F. Evaluation of interactive tactile display system. In *Proc. Int. Conf. on Computers Helping People with Special Needs,* (1998), 29–36.

18. Kelly, D. (1984). Retinal Inhomogenity: I. spatiotemporal contrast sensitivity. *Journal of the Optical Society of America* (1984) 74, 1, 107–113.

19. Lakatos, S. Recognition of complex auditory-spatial patterns. *Perception* (1993), 22, 363–374.

20. Loomis, J. M., and Lederman, S.J. Tactual perception. In *Handbook of Perception and Human Performance,* 2, Cognitive Processes and Performance, Boff, K.R., Kaufman, L., and Thomas, J.P., John Wiley and Sons, New York (1986), Chapter 31.

21. Mannos, J.L., and Sakrison, D.J. (1974). The effects of a visual fidelity criterion on the encoding of images. *IEEE Trans. on Information Theory,* (1974) 20 (4), 525–536.

22. Maucher, T., Meier, K., and Schemmel, J. An interactive tactile graphics display. In *Proc. International Symposium on Signal Processing and its Applications (ISSPA),* Kuala Lumpur, Malaysia, 2001, pp. 190–193.

23. Meijer, P.B.L. An experimental system for auditory image representations. *IEEE Trans. Bio. Eng.*, 39 (2) (1992), 112–121.

24. Parkes, D.N. An audio-tactile device for the acquisition, use and management of spatially distributed information by visually impaired people. In *Symp. on Maps and Graphics for Visually Handicapped People,* (1988) 30–35.

25. Ruff, R.M., and Perret, E. Auditory spatial pattern perception aided by visual choices. *Psychological Research* (1976), 38, 369–377.

26. Way, T.P. and Barner, K.E. Automatic visual to tactile translation, part I: human factors, access methods and image manipulation. *IEEE Trans. Rehabil. Eng.*, 5 (1997), 81–94.

27. World Health Organisation. Magnitude and causes of visual impairment. *Fact Sheet No. 282,* Nov. 2004 (http://www.who.int/mediacentre/factsheets/fs282/en/).

28. Zrenner, E. Will retinal implants restore vision ? *Science 295* (2002), 1022–1025.