Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 9 (2013) 1048 – 1057

Procedia Technology

CENTERIS 2013 - Conference on ENTERprise Information Systems / PRojMAN 2013 - International Conference on Project MANagement / HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies

Color sonification for the visually impaired

Sofia Cavaco[a,*], J. Tomás Henriques[b,*], Michele Mengucci[a], Nuno Correia[a], Francisco Medeiros[c]

[a] *CITI, Departamento de Informática, Faculdade de Ciências e Tecnologia Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*
[b] *SUNY at Buffalo State, 1300 Elmwood Ave, Buffalo, NY 14221, USA*
[b] *Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa Av. Berna 26-C, 1069-061 Lisboa, Portugal*
[c] *LabIO, Rua Nova do Desterro 23 4-Esq., 1100 Lisboa, Portugal*

Abstract

We present a software tool developed to help the visually impaired to perceive characteristics of the environment related to color and luminosity. The tool converts color information extracted from images into sound, where the images can be still images, or video frames from live or stored video. The color information is extracted from the hue, saturation and value attributes, which are mapped into audio attributes directly related to the perception of pitch, timbre and loudness. The tool can be used to gather information about the range of colors present in the images, presence or absence of light sources as well the location and shape of the objects in the images. It is useful for blind people, but it has also applications in other areas such as chemistry and arts.

Keywords: Audio visual system; image color analysis; image representation; image sonification; sensory aids.

* Corresponding author. *E-mail address:* scavaco@fct.unl.pt, henriqjt@buffalostate.edu.

doi:10.1016/j.protcy.2013.12.117

1. Introduction

Sound has a fundamental role in the way we perceive and interact with the environment. It exerts a deep impression on our cognitive system as the neurophysiological ability to decode sounds/vibrations has played a major role in human evolutionary design.

Both sound and light are physical manifestations that can be described in terms of frequency and amplitude, they stem out of periodic oscillations, albeit of different orders of magnitude. Throughout time the link between sound and light has been intuitively understood and explored by scientists and artists. Nowadays we have a body of knowledge that allows us to delve into the intricacies of these two phenomena with much greater insight.

This paper presents a software application that transforms in real time color and light information into sound. It was created to help individuals with visual disabilities to make sense of light and color. The tool is geared at navigation purposes greatly widening a person's cognitive awareness of his/her surroundings. It has applications for sighted individuals too. For instance it can be used by color-blind people to help differentiate colors, and in chemistry laboratories to sonically alert experiments that result in color change. Yet another quality offered within its image-to-sound conversion features is the capability to provide a therapeutic, synaesthetic-like effect, when it maps pleasant images into slow evolving sounds [1].

A digital image is a composition of discrete elementary units, its pixels. The tool converts digital images into sound by mapping the hue, saturation and value attributes of the pixels into audio attributes. The images can be video frames captured in real time, or still images loaded into the application. When a single image or video frame is converted into sound, a composition of elementary sounds originating from the pixels' values takes place. This conglomerate of sonic units carry information about the colors in the image, the amount of darkness/light, the location of the light sources, and the location and shape of the objects in the image.

When the tool captures video in real time, it generates a texture of sound events that pulsates according to a user-controlled scanning rate. Different scanning rates serve different purposes. A high rate is necessary when fast information about the environment is intended (for example, when moving in crowded urban areas). A low rate provides good results to obtain a more detailed description of the surrounding environment.

Other tools that convert image into sound have been developed and reported in the literature. The vOICe, a pioneering work in this area, focuses mainly on scanning shapes but it also provides color information [2-4]. It is quite evolved in terms of the quality of the software and the reach of its distribution. Applications for PC, iPhone and Android are available.

HueMusic is a tool that converts 2D images into music by associating hue values with timbres [5]. Eight different timbres were chosen to represent eight hue values. These timbres/sounds change in amplitude as a function on the presence of each hue in the image. The color composition from still images is conveyed and turned into sound by driving an 8-channel, 8-timbre audio mixer.

Kromophone is a tool that detects color blobs, mapping colors to sound locations [6]. For each blob of a video frame, an averaged area around its center pixels is processed so that the red, green and blue (RGB) parameters within that area are converted into sound. Kromophone only plays one blob/color at a time, translating it into three sounds simultaneously positioned in different locations within the auditory space.

The See Color interface transforms a small portion of a video image into sounds that are represented by musical instruments. These instruments have specific spatial location assignments. See Color quantifies the Hue, Saturation, and Lightness (HSL) color system providing sonification of color and depth, depth being coded by rhythm [7].

Other projects that contain interesting approaches to the image-to-sound conversion and deserve being mentioned are: the Prosthesis Substituting Vision for Audition (PSVA) [8], which maps visual patterns into sound, and the Vibe [9].

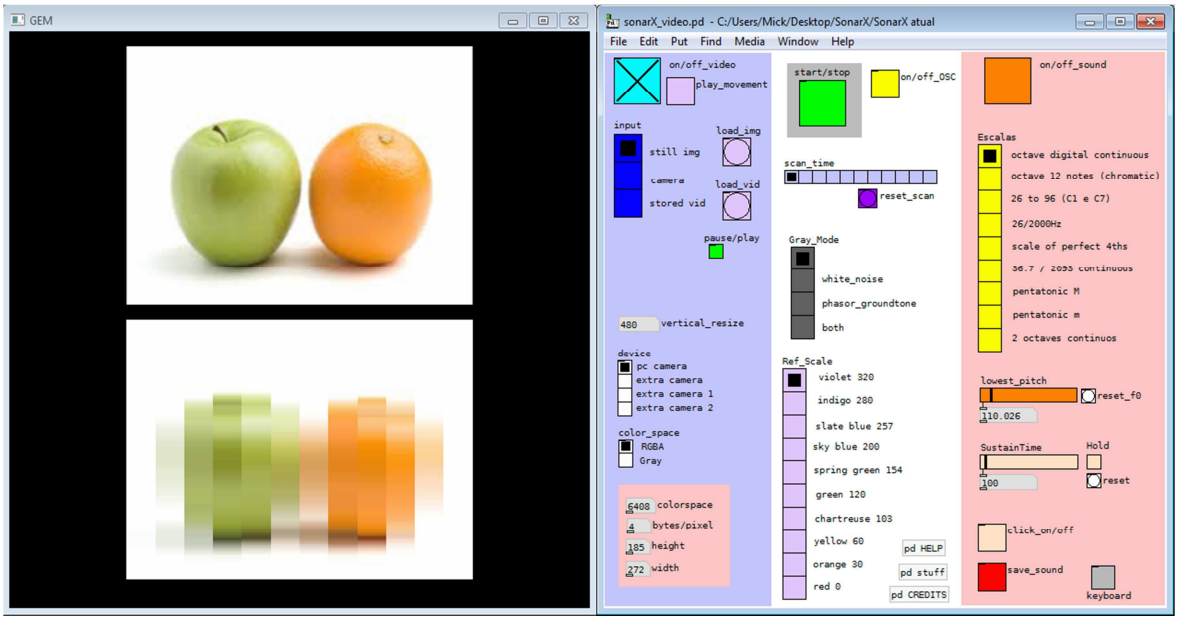

Fig. 1. Interface window.

2. The tool: transforming color into sound

As mentioned above, the tool maps the information in the images (that can be still images or video frames) into audio. Fig. 1 shows the interface window. At the top left it shows the image that has been loaded, and at the bottom left it shows the same image after being processed. The right side shows a series of buttons and sliders that can be used to set several available options, some of which we will discuss below.

The tool was implemented in Pure Data. It has two separate patches that communicate through Open Source Control. Therefore it can run either in one or two machines using a wireless connection. One of the patches processes the video while the other synthesizes the sounds.

The next two sections describe the tool in detail. Section 2.1 starts by explaining how the tool processes the whole image and then section 2.2 discusses how the tool converts the pixels' properties into audio parameters.

2.1. Transforming images into sound

The tool scans the loaded image (or video frame) from top to bottom (Fig. 2) and produces a sound for each row (which, as it will be seen below, consists of the composition of other elementary sounds). The sounds of the rows are played in sequence and not all at once, a process that produces simpler sounds (than the sounds obtained by playing all rows simultaneously) and allows the perception of basic geometric shapes (section 3.3). Other similar tools use this process as well [2, 3]. The scanning rate is user-definable (using a *scan time* ruler, see Fig. 1), and the tool scans and processes the image continuously (from top to bottom), in a loop, until the user decides to stop (*start/stop* button in Fig. 1). The minimum scanning rate is 1 millisecond. This allows the tool to sonify a single frame without sound dropouts.

The tool sections each row in 12 segments of pixels, all with the same length and no overlap. To illustrate this, the bottom left image in Fig. 1, shows the loaded image (displayed on top) after being segmented. The

tool then generates a sound for each of the 12 segments within a row. The 12 sounds are played simultaneously creating a sonic identity associated with each row.

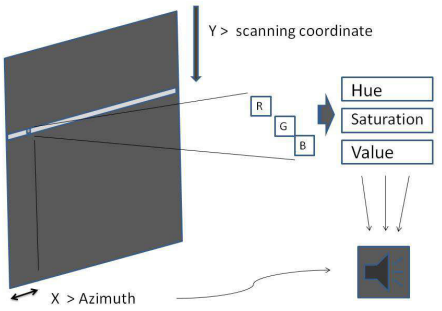

Fig. 2. Diagram of the tool: scanning the image and transforming the HSV pixel parameters into sound attributes.

The segment's center abscissa is used to determine the azimuth of the sound event, which is assigned to one of 12 possible values (Fig. 2). In order to give the perception that a sound event originates from a specific location, we use interaural phase differences: for each segment of pixels, the tool generates a sound and plays that sound over two channels (right and left), where the initial phase of the signals differs in the two channels. The phase difference depends on the segment's center abscissa.

While we use interaural phase differences to give the perception of azimuth, this cue does not give information about elevation. For that we use a click sound (available through the *click on/off* option, Fig. 1) that is played before the sound of the image's top row is played. The time lap between the click and the sound of a shape gives information about the shape's elevation: if the shape is at the top of the image, its sound is played immediately after the click; conversely, if the shape is at the bottom, its sound is played immediately before the click, etc. Consequently, the sounds provide information about the location (azimuth and elevation) of the shapes in the image.

In addition, the generated sounds also give information about motion. If the user pays attention to the differences between consecutive sounds of the whole image, he/she can detect movement of shapes (objects, people, etc.). To facilitate the detection of motion and color changes there is the option of playing only the differences between frames (*play movement* option in Fig. 1). To do this, the tool grabs a snapshot of a scene and subtracts it from all subsequent frames, thus playing only the pixel segments that change.

Another useful available option is *gray mode* (Fig. 1). When this option is turned on, the tool plays the sounds corresponding to gray tones (either as white noise, as the lowest tone in the pitches-span, or as a mixture of both). If the option is turned off, no audible sounds are generated for gray tones. This is useful for pictures with a white background, and for images with white walls (since white is a particular tone of gray, with the *gray mode* option turned off, no sound for the white walls is generated).

2.2. Processing the pixel attribute values

So far we have seen how the tool processes the whole image but we have not seen how it actually generates the sounds, or more specifically, how it transforms color information into sound. For this end, the tool uses the Hue, Saturation and Value (HSV) color model, and it maps the HSV attributes extracted from the image pixels into audio attributes. The HSV color model has been extensively used in computer image analysis as it represents color with attributes closely related to the attributes detected by human vision. Succinctly, hue

contains information about the "pure color" of the pixel, value (or intensity) is proportional to the brightness of the color, and saturation measures the deviation of the color from gray.

For each of the 12 pixels segments mentioned in the previous section, the tool makes an interpolation of the HSV values of the pixels in the segment. The HSV values obtained are converted into sound attributes that give the perception of pitch, timbre and loudness, respectively.

The reason for using these 12 segments is to obtain simpler sounds: instead of playing one sound for each pixel in a row, the tool plays at most 12 sounds simultaneously. For example, with a resolution of 360×480 pixels, the tool treats the image as a matrix of 360 rows per 12 columns (whose values consist of the interpolation of the HSV values of 40 consecutive pixels).

The hue value is mapped into the fundamental frequency, f_0, of the synthesized sound (which gives the perception of pitch). There is an inverse correspondence between the sound's pitch and color frequencies: when the color's frequency decreases from violet to red, the sound's pitch increases (by increasing the f_0) [10, 11]. The synthesized waveform starts off as a sinusoidal wave (i.e., a pure tone), but the final waveform can be different from a pure tone because the signal's spectral envelope can be modified by the other attributes (saturation and value).

The signal's spectral envelope (which is related to the perception of timbre) is controlled by the attribute saturation. The shape of the waveform can vary from a sinusoid (for the lowest saturation value) to a square wave with energy only in the odd frequency partials (for the highest saturation value).

Finally, the attribute value (ranging from 0 to 1) is used to determine the intensity of the signal (which gives the perception of loudness). All frequency partials are affected in the same way as the signal is multiplied by value.

3. Results

3.1. Color

In order to validate the tool we made a pilot study with visually impaired subjects. The goal of this study was to understand if the sounds produced by the tool can efficiently give information about color. There were eight visually impaired volunteers participating in the study, all users or employees from *Biblioteca Nacional de Portugal*. From these eight participants, there were two women and six men with ages between 26 and 63 years old. Most subjects had normal hearing, but two had age related hearing problems.

During the test the subjects were presented 20 sounds and had to establish a correspondence between colors and those sounds. All subjects heard the very same sounds but with different orderings to avoid presentation order effects. The sounds corresponded to the tool's mapping of images that contained only a single color into sound. In this test we used seven colors (red, orange, yellow, green, blue, purple and violet) and the sounds' pitches were all within one musical octave (f_0 varied from 110 Hz to 220 Hz). There was an initial training period of around 10 minutes, so that the subjects could get familiar with the task and the sounds. During the training period the subjects could hear the sounds in the order they wished and repeat them as often as necessary. During training the subjects received feedback about the correct colors that corresponded to the sounds, but once the actual test started no more feedback was given.

The experiment was a forced choice test as the subjects had to pick one of the seven colors for each sound and it was not allowed to have blank answers. Nonetheless, the subjects could take as long as they wished to answer and could hear the sound repeatedly until they felt some confidence to answer.

The overall percentage of correctly classified samples was 51.25% (Fig. 3). In part this may be due to limited training. Subjects reported that they felt that with more training it would be easier to remember the

sound of each color. Indeed, the three subjects that had advanced music knowledge and normal hearing had better results: 68.33% of correctly classified samples (Fig. 3).

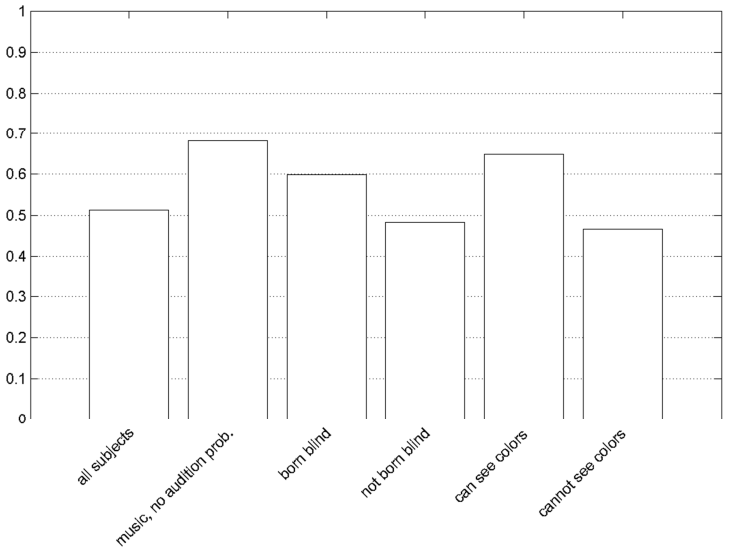

Fig. 3. Percentage of correctly classified samples for different groups: (1st leftmost bar) all subjects, (2nd bar) subjects with music knowledge and no audition problems, (3rd bar) subjects who were born blind or lost vision while still babies, (4th bar) subjects who got blind as older children or even later, (5th bar) subjects who can still see colors or luminosity, (6th bar) subjects with a very high degree of blindness.

The subjects also reported that the pitches-span used in the test was very narrow; with a wider interval it would be easier to memorize and distinguish neighbor sounds. In fact, if we look closely at the results, we can observe that when the subjects give a wrong answer, they pick a color that is in the close neighborhood of the correct color. This can be confirmed by examining the confusion matrix (Fig. 4). The highest values are along the diagonal (except for orange), which correspond to the percentage of correctly classified samples. The positions that are further from the diagonal have lower values. The confusion matrix shows that most wrongly classified samples are mistaken by a color that is close in frequency to the actual color of the sample. For instance, while 71% of the violet samples were correctly classified, 25% of these samples were mistaken for purple and 4.2% were mistaken for blue. This means that while the subjects have no difficulty on distinguishing sounds with very different fundamental frequencies, there is some degree of confusion on distinguishing sounds with neighboring fundamental frequencies.

If we consider color neighborhoods, where a neighborhood of a specific color is that color, the color immediately after it and the color immediately before it ($\{color_{x-1}, \mathbf{color_x}, color_{x+1}\}$), the overall percentage of correctly classified samples (for all subjects) is much higher: 86.25% and varies from 75% for orange and green to 96% for red and violet. Fig. 5 shows these results for the following color neighborhoods: {**red**, orange}, {red, **orange**, yellow}, {orange, **yellow**, green}, {yellow, **green**, blue}, {green, **blue**, purple}, {blue, **purple**, violet}, {purple, **violet**}. For instance, the first white bar corresponds to the neighborhood {**red**, orange} and indicates that 96% of red samples were classified either as red or orange.

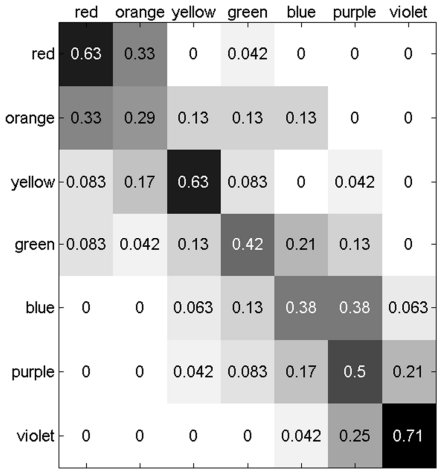

Fig. 4. Confusion matrix for all subjects. The lines correspond to the correct color and the columns correspond to the answers of the subjects. For instance, the 1st line, 2nd column, shows the percentage of red samples that were classified as orange.

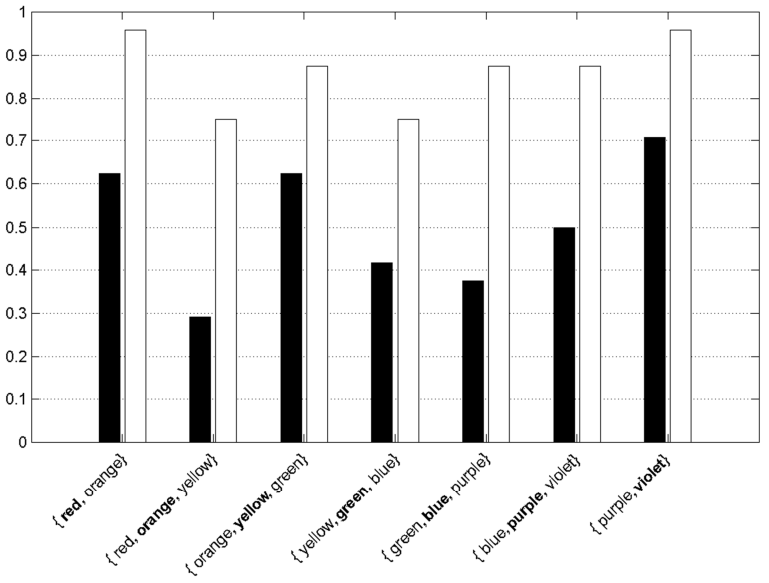

Fig. 5. (Black bars) Percentage of correctly classified samples for all subjects, where the color of the samples are indicated in bold. (White bars) Percentage of correctly classified samples when we consider the color neighborhoods specified between braces.

Fig. 5 also shows the results for all subjects when no color neighborhoods are considered (black bars). It can be observed that the correct classification percentages increase considerably. For instance the first pair of bars indicates that 63% of the red samples are correctly classified but when we take into account the {**red**, orange} neighborhood, this results increases to 96%. These results show that the sounds produced by the tool give the visually impaired enough information about the range of colors in the images.

While so far we have concluded that the tool's mapping of color into sound gives information about the color range in the images, that a wider interval for the pitches is needed and also that more training is needed, there are other observations made during the test that are worth mentioning. Namely, it is interesting to compare the results from people that were born blind with those from people who still remember colors, and the results from people who are still not completely blind (who can still see some colors) with those from the remaining subjects.

While most participants in the study had a very high degree of blindness or were completely blind, three reported that they still could distinguish brightness from darkness or some colors under certain conditions. The overall percentage of correct answers for the later was 65% (Fig. 3) and reached 83.33% for yellow, purple and violet. The remaining subjects had poorer results: their overall percentage of correctly classified samples was 46.67% (Fig. 3). It is interesting to note that for this group, the highest percentages were obtained for red and violet (66.67%), which are the first and last colors in the sequence of colors.

After each subject finished the test, he/she was asked if he/she used any conscious technique to memorize the sounds and the tool's color-sound mapping. Most subjects reported that they used no specific technique. Only two subjects reported having used conscious techniques for this end. These were the only two subjects participating in the study who were either born blind or got blind as still a baby, and these were the only two subjects who did not remember or had never seen color. To them, colors are only a theoretic concept.

The technique they used was to memorize the color order and the sound for one of those colors. One subject memorized the sound of red. The other memorized the sound of green and then tried to understand if the pitches of the samples were above or below the pitch of the sound of green.

Interestingly, the techniques they used proved to be very effective. The overall percentage of correct answers was 60% for the two subjects born blind and 48.33% for the remaining subjects (Fig. 3). It is possible that, since they did not remember colors, they felt the need to use a memorization technique, and, as a consequence, they may have had a more efficient training session which resulted in better results during the actual test session.

3.2. Location

At the present stage of our research, the main goal of the tool is to transform color into sound, but more information can be perceived from the audio it generates. Since the tool uses a click sound to indicate where the scan line is and maps the abscissa of the center pixel of each pixel segment into the azimuth of the sound, it is possible to perceive the location (azimuth and elevation) of the objects in the images.

Apart from the pilot study mentioned above, we also performed other tests with visually impaired people. The goal of one of these tests was to assess if the sounds could be used to distinguish the location of the objects in the images. We loaded images that consisted of a white filled circle on a black background. The circle could be at one of nine different positions: (x,y), where x is top, center or bottom, and y is left, center or right. There was no sound produced for the black background. (The HSV value for black is zero, which is mapped into zero intensity and as a consequence the sound of black is not audible.) The volunteers in this test could say with a high degree of confidence where the circle was (the results were 100% recognition rates for all tested locations).

3.3. Shape

Other information that can be perceived from the sounds generated by the tool is the shape of same-color objects (for simple geometries). Since the tool plays the sounds of the rows in sequence, the way a sound evolves through time, gives information about the shape that is mapped into that sound. For instance, when an

image contains a simple one-color shape on a black or white background (and the *gray mode* option is turned off so that no sound is played for the white background, see section 2.1), the tool produces a sound whose intensity varies according to the shape's geometry. For example, an image with a triangle with the base parallel to the bottom of the interface window, is mapped into a sound that increases in intensity: the sound starts with just one elementary sound because there is only one pixel segment (the segment that contains the top vertex of the triangle) that is mapped into audible sound when the tool processes the first row containing the triangle. As the rows are being scanned (downwards), more and more pixel segments are mapped into sound. The sounds are all equal but as a result the loudness of the whole final sound increases. Therefore, the user will hear a single sound that starts soft and increases in loudness, until the tool scans and processes the row that contains the triangle's base.

It was confirmed by blind volunteers that they could perceive very simple shapes, when the shapes consisted of white filled triangles, rectangles (including squares) and circles on a black background. Yet, these results were only obtained after we explained the blind participants how they could interpret the sounds of the tool: with the help of folded paper models we explained what happens to the sounds while the tool is scanning the shapes.

4. Discussion and conclusions

We presented a software application that converts color information extracted from still images or video frames, into sound. The aim of this tool is to help individuals with visual disabilities in their daily routines, such as finding the exit from a building, crossing roads, choosing clothes with matching colors, detecting motion, etc. With this tool they can gather more information about the surrounding environment. Even though the application has been designed to help blind people, it is also useful for color-blind individuals and it has applications in other areas. It can be used in chemistry laboratories, for instance to signal color changes in chemical reactions. It can be used in visual art exhibitions as well (creating sound from the pieces in the exhibition and thus providing a complement to the visual experience).

The software application uses the HSV attributes of the pixels in the images and maps them into f_0, spectral envelope and intensity, which are closely related to the perception of pitch, timbre and loudness, respectively. It also uses cues to give information of azimuth and elevation. The location of the sound events depends on the location of the shapes in the image.

We performed a pilot study and a set of other less formal tests with blind participants in order to validate the tool. These studies showed that the synthesized sounds give information about the pixels' color range, the location of the shapes, and the geometry of one-color, simple shapes. Other information that can be obtained from the sounds includes: number of objects in the scene, location of light sources and detection of moving shapes.

The results discussed in section 3.1 suggest that representing seven colors within one musical octave is not enough; having a wider pitch interval between the different sounds would be a better option. As a response to this observation, the application now includes the option of working with wider pitches spans (the *scale* options in Fig. 1). Now it is possible to map the sounds into one octave, two octaves, a pentatonic scale, among other options. It is also possible to map the colors into a small discrete number of sounds (in which case there will be quantization; similar colors, such as different shades of green, may be mapped into the same sound) or to use a (closer to) continuous set of tones (in reality this is just a very high discrete number of sounds). The lowest tone of the pitch space is also user adjustable (option *lowest pitch* in Fig. 1).

Even though further tests need to be done, we are convinced that the order of the color-to-sound correspondence is important. That is, that the fundamental frequencies of synthesized sounds must preserve the natural color order (at least for people who were not born blind and remember colors). The tool preserves

this ordering but it uses an inverse correspondence between the sounds' pitches and color frequencies (section 2.2). An alternative would be to invert this ordering. One participant in the tests suggested that having an inverse association of colors to pitches would be more intuitive as he would rather have warm colors (like red) associated to the lower pitches.

As mentioned above, the amount of training is important and affects the results of the tests. In order to facilitate the learning process of associating sounds to colors, the application now includes the option of playing 10 reference sounds (*ref scale* in Fig. 1). When the user selects any of the reference colors, he/she will hear a voice saying the name of the color and then the sound of that color.

As future work we plan to make the tool easier to manipulate by blind users. Many of the options can already be chosen through the keyboard. Yet, there is still difficulty manipulating the video camera as blind people tend to move it too fast. Also presently the tool runs on a standard laptop connected to a camera. We are currently exploring a couple of extensions to this work that are easier to manipulate and will allow greater mobility for the user. On the one hand, a hardware prototype is currently being developed. It will add distance and orientation sensors as well as vibrotactile feedback to obtain greater insight about the surroundings and to decode further information within the images. On the other hand, we are also exploring other possibilities that can be easily acquired by the blind population, such as a mobile application that runs on smartphones. Finally, the tool will also include other image processing techniques such as edge detection, face detection and object tracking.

Acknowledgements

This research was supported by grant-Exp/MAI/0025/2009 from the Portuguese Foundation for Science and Technology and the UT Austin Portugal Program in Digital Media.

The authors thank all participants in the two user studies. Special thanks to Carlos Ferreira for his collaboration and help provided setting up the user studies.

References

[1] M. Mengucci, J.T. Henriques, S. Cavaco, N. Correia, and F. Medeiros, "From color to sound: Assessing the surrounding environment," in Proceedings of the Conference on Digital Arts and New Media (ARTECH), 2012, pp. 345–348.

[2] P.Meijer, "An Experimental System for Auditory Image Representations",. IEEE Transactions Biomedical Engineering, 1992, vol. 39, pp. 112-121.

[3] M. Auvray, S. Hanneton, and J.K. O'Regan, "Learning to perceive with a visuo-auditory substitution system: Localization and object recognition with the vOICe," in Perception, 2007, vol. 36(3), pp. 416–430.

[4] P. Bach-y-Rita and S.W. Kercel, "Sensory substitution and the human-machine interface," in Trends in Cognitive Neuroscience, 2003, vol. 7(12), pp. 541–546.

[5] D. Payling, S. Mills and T. Howle, "HueMusic – creating timbral soundscapes from colored pictures", in Proceedings the of International Conference on Auditory Displays (ICAD), 2007.

[6] Z. Capalbo and B. Glenney, "Hearing color: Radical pluralistic realism and SSDs," in Proceedings of AP-CAP, 2009, pp. 135–140.

[7] G. Bologna, B. Deville and T. Pun, "Sonification of Color and Depth in a Mobility Aid for Blind People", in Proceedings of the International Conference on Auditory Displays (ICAD), 2010.

[8] C. Capelle, C. Trullemans, P. Arno, and C. Veraart, "A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution," in IEEE Transactions Biomedical Engineering, 1988, vol. 45, pp. 1279–1293.

[9] M. Auvray, S. Hanneton., C. Lenay, and K. O'Regan, "There is something out there: distal attribution in sensory substitution, twenty years later," in Journal of Integrative Neuroscience, 2005, vol. 4, pp. 505–521.

[10] D.L. Datteri and J.N. Howard, "The sound of color," in Proceedings of the International Conference on Music Perception and Cognition (ICMPC), 2008, pp. 767–771.

[11] I.C. Firth, "On the linkage of musical keys to colours," in Speculations in Science and Technology, 1981, vol. 4, pp. 501–508.