From Color to Sound: Assessing the Surrounding Environment

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

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

mengucci@gmail.com, henriqjt@buffalostate.edu, scavaco@fct.unl.pt, nmc@fct.unl.pt

Abstract - **A prototype that converts color information of an image into sound, mapping color attributes to sound parameters (pitch, timbre, volume, localization) is described. The sonic print of an image is created as a composition of sounds originating from its elementary components, the pixels.**

By converting a stream of video frames into sound, the tool produces a complex and pulsating texture of sonic events. The tool aims to aid the visually impaired to infer the world of light and color, and also anyone who needs to obtain information from video without direct visual access. It can also be used for artistic and therapeutic purposes.

Index Terms - **Audio visual systems, computer application, data conversion, image color analysis, image representation, sensory aids.**

I. Introduction

Sound has a fundamental role in the perception of the world, as it is strongly connected to simple principles of physics that represent and define nature. Both sound and color are physical phenomena that can be expressed and measured in terms of their frequencies. Throughout the history of mankind, a bond between color/light and sound has inspired the work of artists and scientists alike. From Pythagoras[1], and Aristotle [2] to Newton [3], Scriabin [4] and Kandinsky [5], many have been seduced by this mysterious link. Nowadays technology allows a much deeper interlacing of these two fields of human perception [6]-[12].

This paper presents a simple and robust digital prototype that transforms, in real time, color and light information into sound. This prototype was created to help individuals with visual disabilities make sense of light and color. It can specifically be utilized for navigation purposes or more generally to widen their cognitive awareness. The tool is also useful for those who need to obtain information about the surroundings without directly looking at them. It can be further used as a complement to artistic exhibitions (architecture, sculpture, painting, performance) and for therapeutic purposes.

A digital image is a composition of discrete elementary units of color/light, its pixels. The main idea of our research is to convert a digital image into its sonic print, where each pixel, pixel properties or group of pixels have an associated sound. Therefore, when a single image or video frame is "played", the result is a composition of elementary sounds originated directly by the pixels' values. The sound events, generated while the images are scanned, create sounds of varying complexity that convey what the camera captures, giving information about: the colors that surround the user, the amount of darkness/light, and the location of these light sources.

Applied to live video streaming, this prototype generates a pulsating texture of sonic events. The scanning rate is user controlled to adapt to different needs: a high rate when fast information about the environment is necessary (for example, when moving in crowded urban areas) and a low rate if a detailed description of a scene is desired.

Other instruments have been created to convert video to sound for similar purposes. These include scanning shapes, like the vOICe [7]-[8], making use of the translation of visual patterns into sound, like the PSVA (Prosthesis Substituting Vision for Audition) [9], finding virtual "sources" in the image and outputting a mixture of sinusoidal waves, like the Vibe [10] and the Kromophone [11], which maps colors to sound locations on the auditory field.

II. Overview of the Prototype

Throughout time there have been several attempts to establish a direct relationship between color and sound [3]-[4]. Both phenomena are manifestations of energy, measured in terms of their oscillatory fundamental frequency (f_0). Sound covers a range of frequencies that span ~20Hz to ~20000Hz. The visible range of colors goes from red ~405-480 THz to violet ~700-790 THz. We chose to transpose the colors of the visual spectrum into the auditory range by slowing down their

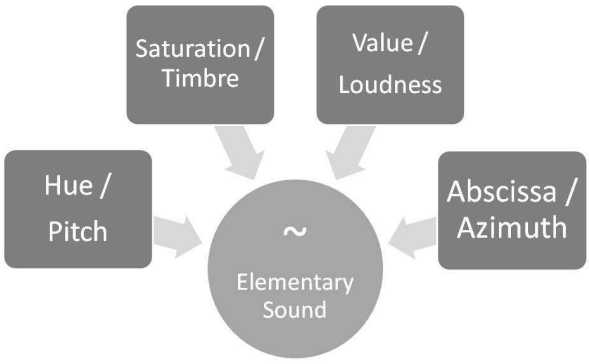

Fig. 1. Elementary sound. Relationship between color and light parameters and sound parameters.

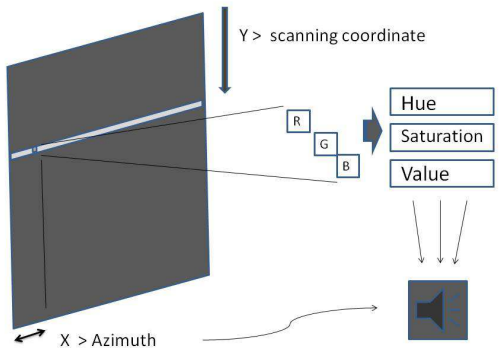

Fig. 2. Diagram of the prototype.

frequencies, such that we got frequencies that approximately span the interval of one musical octave, in the mid-range. (Two frequencies separated by an octave have a ratio of 2:1, that is, the highest frequency in the interval is twice as large as the lowest frequency) [13]. The key idea we used to map color properties into sound was thus to create a tool that assigns and plays a different tone for each color either within one or two musical octaves.

Below we will see how we map the color information into sound in more detail (section II.A) and how we process the whole image (section II.B). Section II.C includes a description of the technical details.

A. Mapping color to sound

As mentioned above, the tool converts the color information present in the image into sound. For that purpose we use the Hue, Saturation and Value (HSV) color model. This model is widely used in image analysis as it represents color in a way that is closer to our perception when compared with the RGB color model. Hue contains information about the "pure color" of the pixel. Value, sometimes called intensity is proportional to the brightness of the color. The third descriptor, saturation, measures the deviation of the color from gray.

As illustrated in fig. 1, the prototype maps the color HSV attributes of the pixels into sound parameters respectively as: the f_0 (which is related to the sound's perceived pitch), the spectral envelope of the sound (which influences the perception of timbre), and the intensity (which is related to the sound's perceived loudness). The prototype also uses a forth parameter, the pixel's abscissa, which is mapped into azimuth, that is the simulated position of the sound source in the horizontal plane.

The sound output of a hue value is a sinusoid whose frequency depends on this value. The sound output of an HSV value is a sound whose f_0 belongs to the interval of one or two octaves, divisible into a number of tones that depends on the camera's color resolution. The lower frequency colors, like red and orange, give a sensation of strength and power, relating better to higher pitched sounds. Therefore, the mapping used inverts the association of color and sound frequencies, that is, when the color's frequency decreases from violet to red, the sound's pitch increases (by increasing its f_0). This respects how most humans feel the sound of the colors [14]. The lowest tone of the pitch space is user adjustable.

The saturation value is used to modify the timbre of the sound. This value is used on a waveshaping function that, using a clipping distortion, gradually modifies the shape of the waveform from a plain sinusoid (corresponding to the lowest saturation value) to a square wave with energy in the odd frequency partials (corresponding to the highest saturation value). As a consequence, the spectral envelope and, consequently, the timbre of the sound change.

The intensity of the sound is controlled by the third parameter, value. All frequency partials are affected in the same way as the sound is multiplied by value (whose range is from 0 to 1).

B. Processing the whole image

The rate at which the video frames are scanned is user controlled. In fast mode it allows the user to detect fast changes in the environment and to quickly locate the color light sources. At a slower mode it permits the user to appreciate the details of the scene and objects captured by the camera. In addition, the tool allows the user to grab a snapshot of a scene and subtract it from all subsequent frames, thus playing only the pixels that change. This is a very useful feature for detecting motion and color changes on a fixed background.

Our first image-to-sound conversion approach was to reduce the image into homogeneous regions, such as regions that corresponded to blobs with the same color and texture. However playing the image in its basic form as a simple grid of pixels, with each pixel being a light/color source that hits the eye to form shapes and give meaning to external objects in our brain, proved to

be a more effective and robust technique. This method sonically emulates the impact that light has in our vision, and provides a platform onto which higher-level techniques can be added.

The tool scans the images from top to bottom. Instead of generating an elementary sound for each scanned pixel, the tool makes an interpolation of the HSV values over 12 segments along the rows and generates a sound for each of those segments (fig. 2). Finally, the segment's center abscissa is used to determine the azimuth of the simulated sound source, which is assigned to one of 12 possible values.

Like with other similar devices [7]-[8], the tool plays the image as it is being scanned, and not all at once. As a result, the tool interprets the raw color/light information and sends out a simple but robust signal.

C. Technical details

The software is implemented in Pure Data with two separate patches, one for video processing and the other for sound synthesis. These patches communicate through Open Sound Control and can run in one or two separate machines linked with a wireless connection. The image processing part of the tool runs with PD extension GEM (Graphics Environment for Multimedia)[1]. The tool is currently being adapted in order to run with MaxMSP. We used a Playstation PS3 Eye camera, which was chosen for its robustness and ability to capture standard video with frame rates of 60Hz at a 640×480 pixel resolution. The processing time for a video frame with this resolution is 0.192 seconds, which gives a rate of 5.208 loops per second but it can be lowered with less vertical resolution.

III. USAGE

In order to tune all the parameters (the range of notes that corresponds to hue, how the timbre depends on the saturation, etc.) the tool has been tested with both sighted and visually impaired subjects. Here we describe the conclusions that we drew from these tests.

Subjects were encouraged to practice with the tool to become familiar with the sounds generated and the information it conveys. The range for f_0 was from 220 Hz to 440 Hz (notes A3 to A4) and from 220Hz to 880Hz, respectively for the one octave and two octaves.

A. Tests and Applications: visually impaired subjects

A preliminary test was performed by a visually impaired (musician) subject who reported great interest and satisfaction. This test was performed indoors, in the subject's own environment. Later we performed a formal test with 8 visually impaired participants at Biblioteca Nacional de Portugal. At a preliminary contact, the tool allowed immediately the recognition of light and colors and the subjects suggested a number of possible utilizations to infer daily routines details, largely underestimated by the majority of people. For example, the tool has proven to be valuable for choosing pleasant color combinations of clothes to wear, which is a common dilemma for blind people. The tool can also help to acquire knowledge about the weather: pointing the camera device at the sky, the user can infer if it is cloudy (gray) or sunny (blue). In indoor environments it can help locating the exit from a public building, like a bank or post office (by emitting a loud sound, associated to a strong light). Also common daily life objects, such as doors, tables, or restrooms, can be readily identified.

B. Tests and Applications: sighted subjects

The tool is not only useful for the visually impaired, but it may also be used by sighted people to augment one's awareness of the surroundings that cannot directly be seen. For instance, it can let the user know if the camera is pointing to a texture (homogeneous sound) or to an object (some particular sound event happening during the scan loop). It can also be useful as a new type of motion detector that gives out relevant information about what type of object/person/etc has been captured.

Finally, it can be used as a therapy tool if played in slow mode and with "pleasant" images. The synesthetic effect of looking at a colorful image while listening continuously to the sound it creates can be very relaxing and therapeutic according to the combination of sound and color therapy principles. This effect has been experienced during a test of the tool at a Mandalas drawing workshop. (Mandalas are symbols used in Buddhism and Hinduism. Drawing Mandalas is commonly used for relaxation and stress reduction purposes.) This test was performed outdoors.

After listening to the "sound of the sky" (by pointing the camera to the sky), subjects pointed the camera to the pictures that they drew and painted during the workshop, and listened to the music created by the prototype through headphones. In some cases the sound did not match the expectations, in others the users got really excited and pleased, and kept on listening to the articulated sound. It was noted that the generated sounds were capable of inducing deep relaxation effects.

IV. RELATED WORK

There are other software tools geared at helping the visually impaired to infer the world by mapping visual data into auditory information. These tools and devices, like ours, attempt to provide synthetic vision to the user by means of a non-invasive visual prosthesis. One of the most known of such tools is vOICe [7]-[8]**.** This software scans the image from left to right and combines a threshold detection technique with an association of height to pitch and brightness to loudness, to express

[1] http://www.danks.org/mark/gem/

sonically the shapes in the images. Views are typically refreshed about once per second. The software makes use of high level semantics and sophisticated tools to describe the images. Nonetheless it does not take into account light and color, focusing just on shapes. While shape is perceptible to a blind person, through touch, it is impossible for him/her to perceive light and color. Moreover, once these are well described and located, the information about shape becomes attainable.

The Vibe, an open source project hosted by Sourceforge [10], produces sound as a mixture of sinusoidal waveforms. These are created by virtual "sources", each corresponding to a "receptive field" in the image. The Vibe has an approach similar to ours but it is more processing intensive since it needs to find *regions* in the image.

Another interesting and more recent project in this area is the Kromophone [11]. Like our tool it focuses on colors rather than shapes. It distinguishes colors based on their RGB attributes, and translates that information mainly into sound localization cues, more so than pitch or timbre.

Another successful visual-to-auditory sensory substitution device is the Prosthesis Substituting Vision for Audition (PSVA) [9]-[10]. This device utilizes a head-mounted TV and translates visual patterns into auditory patterns, with a system that uses pixel to frequency relationship and couples a rough model of the human retina with an inverse model of the cochlea.

V. Future Research and Conclusion

A prototype that converts image into a sonic print was created to help individuals with visual disabilities. The key idea of this tool is converting the light and color information (coded with HSV attributes) into sound parameters: f_0, intensity and spectral envelope (which are related to the perceptual measures of pitch, loudness and timbre, respectively).

Though the prototype is especially aimed at helping visually impaired individuals in their daily routines, such as finding the exit from a building or crossing roads, it can be useful to everybody. It allows to infer the information of a scene captured by a camera through sound, and also has other functionalities such as artistic and therapeutic.

From the arts and performance perspective, the usefulness of the tool is noteworthy, as it lends sound to images giving them a new dimension. The tool has also been successfully applied to interpret relaxation drawings. To further this purpose we mapped color into larger pitch spans used particular musical scales as well.

As future work we plan to add image processing techniques such as edge detection, face detection and object tracking. The output of these processes will be mapped to more complex sounds as well as rhythmic patterns.

Acknowledgements

This research is made possible through the support of the Portuguese Foundation for Science and Technology (grant UTA-Exp/MAI/0025/2009) and the UT Austin Portugal Program in Digital Media.

The authors also wish to thank Sebastião Antunes and Rui Jesus for the help provided through the development of the prototype and João Silva allowing testing the prototype in the relaxation drawing workshop. Many thanks also to all participants in the test and to Carlos Ferreira from Biblioteca Nacional de Portugal for all his interest, feedback and for the help provided setting up the test performed at that institution.

References

[1] E.G.McClain, *The Pythagorean*, Plato.Stony Brook, NY.: Nicolas Hays, 1978.

[2] Aristotle, “Sense and sensibilia”, *The Complete Works of Aristotle*, vol. 1, pp 693-713, Princeton, NJ: Princeton University Press, 1984.

[3] I.Newton, *Opticks or, a treatise of the reflexions, refractions, inflexions and colours of light: also two treatises of the species and magnitude of curvilinear figures*. Commentary by Nicholas Humez (Octavo ed.). Palo Alto, Calif.: Octavo, 1998.

[4] M.D.Garcia, E.Emanuel, "Scriabin's Mysterium and the Birth of Genius", *Mid-Winter Meeting of the American Psychoanalytic Association*. New York, New York. Retrieved, 2007.

[5] M.Henry: “Seeing the Invisible”, *On Kandinsky* Continuum, 2009.

[6] P.B.L.Meijer, “An Experimental System for Auditory Image Representations”,. *IEEE Transactions Biomedical Engineering*, vol. 39, pp. 112-121, 1992

[7] M. Auvray, S. Hanneton and J.K. O'Regan, "Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with ‘The vOICe’", *Perception*, 36(3), pp. 416-430, 2007.

[8] C. Capelle, C. Trullemans, P. Arno, C. Veraart, "A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution.". *IEEE Transactions Biomedical Engineering*, vol. 45, 1279-1293, 1998.

[9] P. Bach-y-Rita, S.W. Kercel, “Sensory substitution and the human-machine interface”, *Trends in Cognitive Neuroscience*, 7(12):541-546, 2003

[10] M. Auvray, S. Hanneton., C. Lenay, K. O'Regan, “There is something out there: distal attribution in sensory substitution, twenty years later”, *Journal of Integrative Neuroscience*, vol. 4, pp. 505-21, 2005.

[11] Z. Capalbo, B. Glenney, “Hearing Color: Radical Pluralistic Realism and SSDs”, *Proceedings of AP-CAP*, pp. 135-140, Tokyo, Japan, October 2009.

[12] N. Harbisson, "Painting by ear" Modern *Painters, The International Contemporary Art Magazine*, pp.70-73, New York, June 2008.

[13] I.C. Firth, “On the linkage of musical keys to colours”, Speculations in Science and Technology, Vol 4, 501-508, 1981.

[14] D.L. Datteri and J.N. Howard, (2004) “The Sound of Color”, Proceedings of ICMPC, pp. 767-771, 2004.