FROM LIGHT TO SOUND: SIGNAL MODALITY TRANSLATION

Jiaxuan Wang School of Electrical and Electronic Engineering, University of Manchester, UK
Patrick Gaydecki School of Electrical and Electronic Engineering, University of Manchester, UK

1 INTRODUCTION

This paper describes a signal modality translation project the purpose of which is to represent acoustic signals from images or patterns of light. Signal modality translation is a rapidly emerging technology; it has been widely used to synthesize images from wave signals or acoustic signals. Conversely in this project, signals in the form of images and video are modulated into continuous acoustic sounds, by the method of real-time digital signal processing (DSP).

The hardware equipment of this project is a conventional CCD camera, signal processing core, and an audio output system. The software equipment is MATLAB software with the Image Acquisition Toolbox.

The whole structure of the project comprises four process stages: the Image Acquisition System, the Feature Extraction System, the Sound Modulation System and the Sound Playback System. The output of this entire system is acoustic audio composed of various instrument library tones and is modulated according to the image features of the video input. Each audio output is not only unique as the video content changes, but also features continuous tones with distinct combinations of pitches and timbres.

Figure 1: Four main stages in this project.

This system could bring benefits to individuals with visual impairment, by aiding them in navigation and location. Moreover, it could be developed to applications in recognition systems, because visual information in some circumstance is difficult to extract.

2 SYSTEM CONTENT

2.1 Image acquisition process

As this system is based on images for acoustic modulation, the first procedure is to acquire images from the external environment, which is called the Image Acquisition Process. Both hardware and software tools are required. For the hardware a Logitech HD Webcam C310 was used as the imaging front-end, while for the software, a MATLAB (R2014b) program with an Image Acquisition Toolbox (R2014b) was employed.

In this process, the webcam was used to capture real-time figures with an option for an adjustable frame rate. This provided flexibility for real-time calculation of statistics and fluency in acoustic modulation. Each figure was imported to Matlab via the Image Acquisition Toolbox, while it was also saved as a temporary backup file in the computer. The specifications of the webcam are shown in Table 1.

General Settings	FrameGrabIntercal	1
	FramesPerTrigger	10
	Logging Mode	Memory
	Name	RGB24_640x480-winvideo-1
	NumberOfBands	3
	ROIPosition	[0 0 640 480]
	Timeout	10
	Type	videoinput
	VideoFormat	RGB24_640x480
Colour Space Settings	BayerSensorAlignment	GRBG
	ReturnedColorSpace	RGB
Acquisition Sources	Source	[1x1 video source]

Table 1: The property of acquisition device

An image is a rectangular array of values (pixels). Over a finite area, each pixel represents the measurement of some properties of a scene, characterized by a fixed number of bits. For images with colours, each pixel is composed of three layers, represented as the red(R), green (G), and blue (B) channels respectively. The features of an image comprise many aspects, while the most usually measured feature are the brightness of a colour image or in RGB channels respectively.

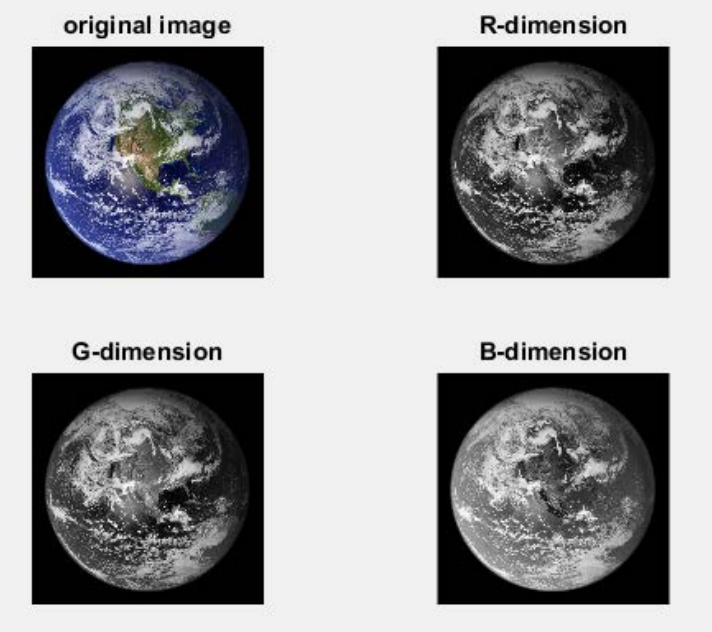

Figure 2: The red, green and blue dimensions of an example image.

2.2 Feature Extraction System

After the images are collected and stored, they are analyzed into digital parameters. First, Fast Fourier Transform is applied in the analysis. The Fast Fourier Transform (FFT) is an efficient implementation of the Discrete Fourier Transform (DFT) used in digital image processing [1]. This algorithm is applied to convert an image from the spatial domain to the frequency domain [3].

Technically speaking, the Fast Fourier Transform is an algorithm and the Fourier transform is the mathematical transformation that it computes.

The FFT requires that the dimensions of the image are a power of two (2^N) in the application. Another characteristic of the FFT algorithm is that the transform of N points can be rewritten as the sum of two N/2 transforms (divide and conquer) [4]. This is the key point since some of the computations could be reused to eliminate redundant operations.

The calculation of the Fourier Transform results in a set of complex numbers for the image in the spatial frequency domain. So they are stored as floats, in order to keep these values precise. Additionally, since the dynamic range of the Fourier coefficients is too large to be displayed on the screen, the values are scaled to an appropriate range. The scaling method is usually dividing them by Height * Width of the image [3].

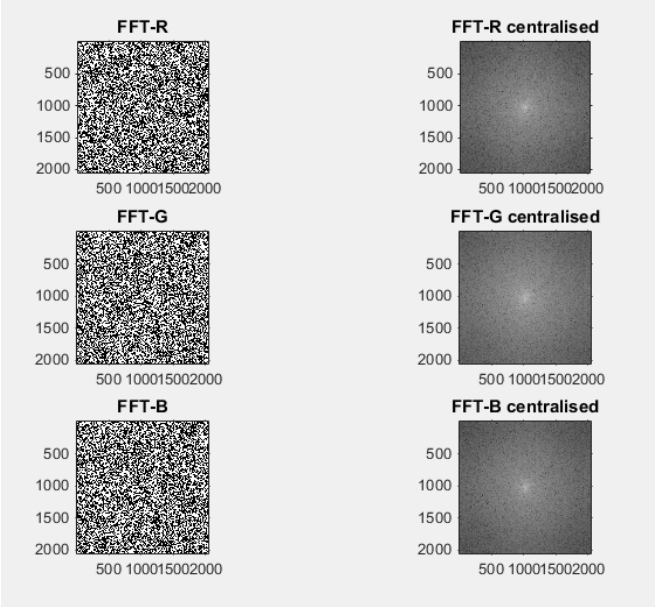

Figure 3: the Fast Fourier Transform (FFT) result of the example image in figure 2.

Image feature extraction methods are useful tools for distinguishing objects in images as well as matching different image segments. Thirteen parameters in total were used and they are composed of two types: spatial features and frequency features. Especially for the frequency features, each of them is the mean value or variance of three values that are respectively calculated in the red, green and blue (RGB) channels. Table 2 lists the spatial features and frequency features used in this system.

Spatial Feature	Frequency Feature
Average Grey Value	Spectrum Mean Energy
Average Contrast	Maximum Amplitude
Smoothness	Frequency of Maximum Amplitude
Third Moment	Minimum Amplitude
Uniformity	Frequency of Minimum Amplitude
Entropy	Maximum Frequency
	Minimum Frequency

Table 2: The features used by the Feature Extraction System

2.3 Sound Modulation System

The Sound Modulation System and Sound Playback System are as shown in Figure 4.

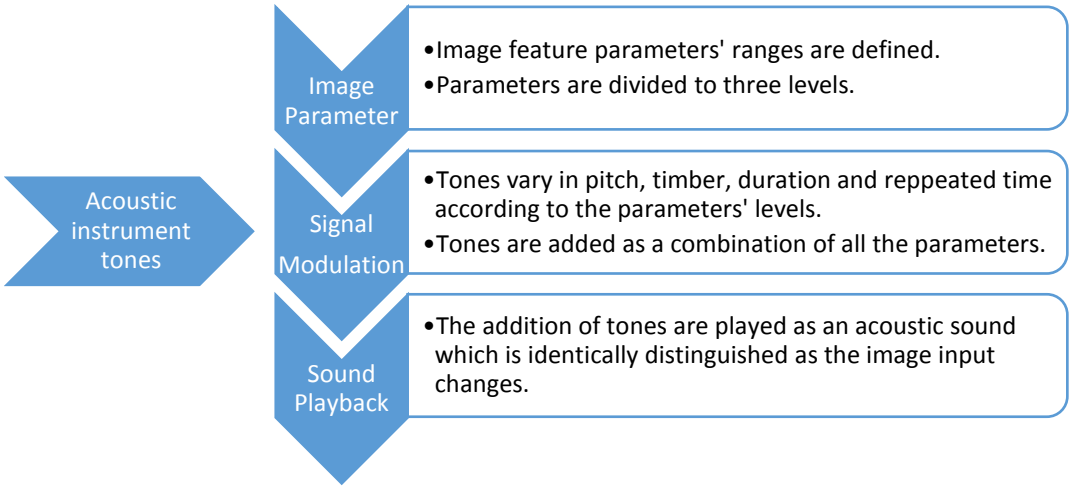

Figure 4: Modulation method in Sound Modulation System.

2.3.1 Image parameters

In the sound modulation system, each parameter obtained from the previous step is statistically analyzed and assigned a category or level. This allows for highly resolved discrimination of the input image. In this system, all thirteen parameters are assorted into three levels, so there are 39 statistical levels in total.

Parameters	Lower Level	Medium Level	Higher Level
Average Grey Value	< 110	$100 - 150$	> 150
Average Contrast	< 40	$40 - 70$	> 70
Average Smoothness	< 0.04	$0.04 - 0.08$	> 0.08
Average Third Moment	< 0	$0 - 1.5$	> 1.5
Third Moment Variance	$< 50 * 10^{-3}$	$50 - 300 * 10^{-3}$	$> 300 * 10^{-3}$
Average Uniformity	$< 70 * 10^{-4}$	$70 - 100 * 10^{-4}$	$> 100 * 10^{-4}$
Average Entropy	$< 700 * 10^{-2}$	$700 - 750 * 10^{-2}$	$> 750 * 10^{-2}$
Average Maximum Frequency	$< 60 * 10^{5}$	$60 - 75 * 10^{5}$	$> 75 * 10^{5}$
Average Minimum Frequency	< -1	$-1 - 0$	> 0
Average Maximum Amplitude	$< 130 * 10^{5}$	$130 - 150 * 10^{5}$	$> 150 * 10^{5}$
Average Minimum Amplitude	< 3	$3 - 5$	> 5
Average Mean Energy	$< 80 * 10^{7}$	$80 - 100 * 10^{7}$	$> 100 * 10^{7}$
Mean Energy Variance	$< 1 * 10^{3}$	$1 * 10^{3} - 3 * 10^{3}$	$> 3 * 10^{3}$

Table 3: Image feature parameters and their levels

2.3.2 Acoustic instrument tones

On the basis of the tone library, diverse combinations of output tones can be defined and attached to permutations of levels in the calculated parameters. As the result of the modulation process, tones with distinct pitches, timbres, length and repeat times are produced, and the tones gradually alter as the camera moves over different scenes. The sound library includes tones from three

instruments: piano, cello and violin, each of which has a distinctive timbre. While for each timbre there are 1/4 notes, 1/8 notes, and 1/16 notes, moreover there are 3 scales on tones for each note.

2.3.3 Modulation method

Modulation is the process of conveying a message signal inside another signal that can be physically transmitted. It varies one or more properties of a carrier signal with a modulating signal. The carrier signal is a high-frequency periodic signal, while the modulating signal typically contains information to be transmitted. Both frequency modulation (FM) and amplitude modulation (AM) are the general types of modulation method.

Amplitude modulation is mainly used in the Sound Modulation System described here. By modulating the time variables using AM, a library of continuous scales of tones is created as the tone source. In the library, the timbre of tones may be chosen as cello, piano, and violin.

On the basis of the tone library, diverse combinations of output tones are defined and attached to permutations of levels in the calculated parameters. Hence tones with distinct pitches, timbres, lengths and repeat times are produced when any parameters changes. For example, the average grey value has a direct influence on the length of the result. As is listed in table 3, there are lower level, medium lever and higher level for the average grey level value, and they respectively represent the 1/16, 1/8, and 1/4 notes of the tone which determines the length of tone. Consequently, each time (frame) the webcam captures an image, the average grey value would be calculated in the Feature Extraction System, therefore in the following Sound Modulation System, the program will find out the parameter level for the average gray value, then produce the corresponding length of tones.

2.4 Sound Playback System

The Sound Playback System is the final component of the project. It is defined as a system that can play the acoustic sound which is modulated in the previous Sound Modulation System. In this system, the sound is played with sensitive and apparent variation in the pitch, length, timbre and repeated times of each tone. Thus when the image in front of the camera is changed on a small scale, the resulting sound is different in a readily identifiable way.

3 RESULTS

3.1 Image Acquisition and Feature Extraction System Result

Figure 5 shows the screenshot of the real-time Image Acquisition System and Feature Extraction System. To start with, the real-time video is shown in the lower-right of the screen. The captured image is shown in the upper-right part of the screen, in which the frame rate may be set according to Table 2. The MATLAB program is shown in the left part of the screen, and displays and updates all the parameter of the captured image.

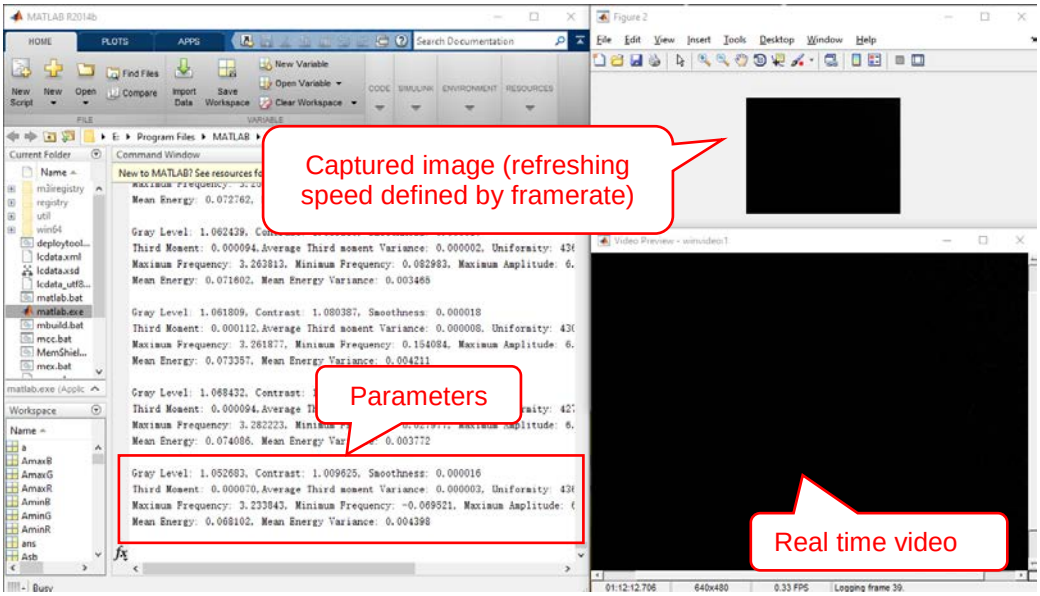

Figure 5: The screenshot of real-time image acquisition and feature extraction step.

3.2 Sound Modulation System Result

In Figures 6 to 9, four screenshots of the system output are shown in response to various scenes imaged by the camera. It may be observed both from the real-time acoustic sound and from the output signal waveform, that as the captured image changes, the output sound changes in different aspects, such as the repeat time, pitch and length. Moreover, this system is sensitive to any small movement in front of the camera, which is translated as a change in tonal quality.

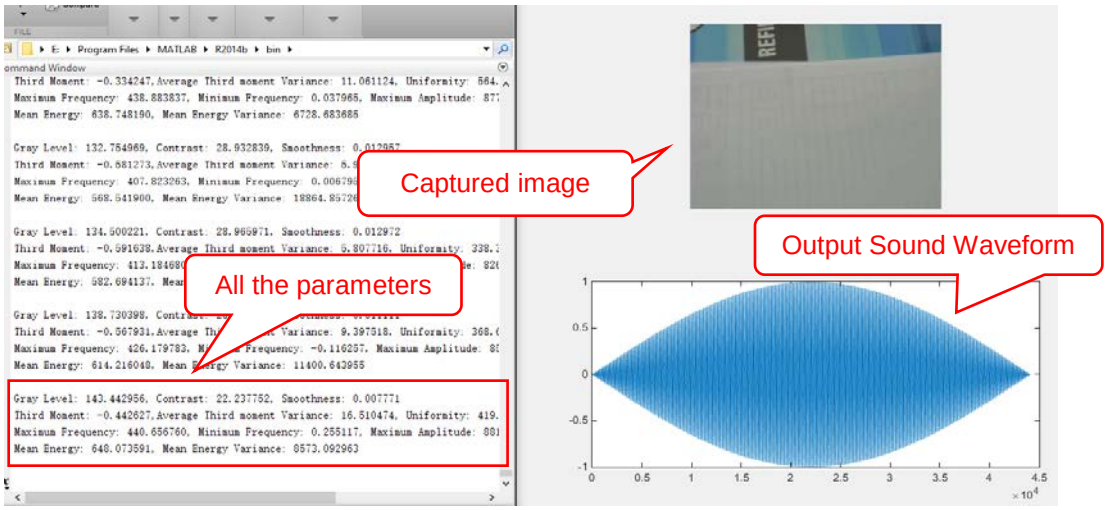

Figure 6: The screenshot of system result 1.

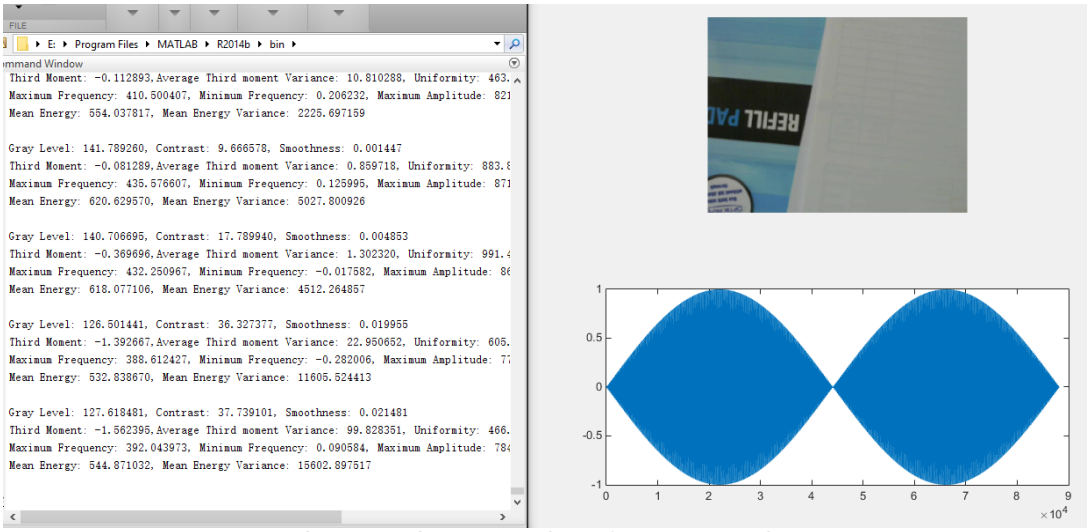

Figure 7: The screenshot of system result 2

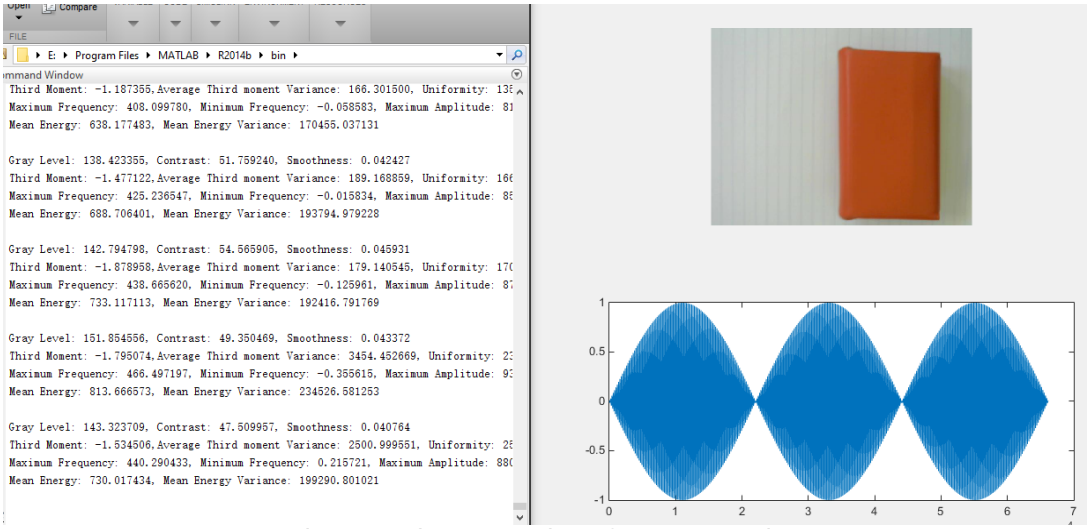

Figure 8: The screenshot of system result 3

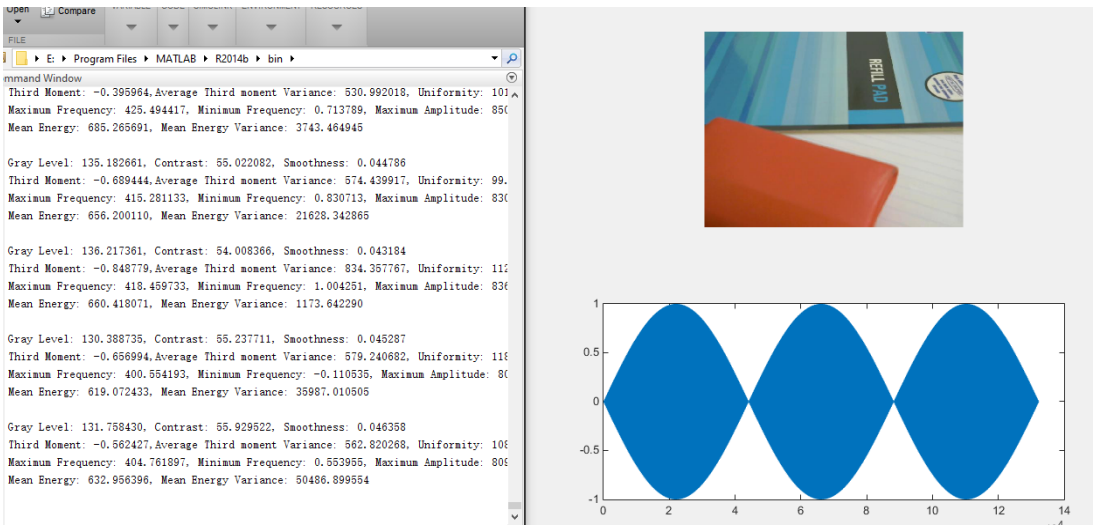

Figure 9: The screenshot of system result 4

4 CONCLUSION AND FURTHER DEVELOPMENT

This system successfully achieves the translation from image information to simple acoustic sound signals. The algorithmic structure of the system is largely complete and the most important modulation methods have been developed. The outputs are uniquely associated with the given inputs. However, there is still considerable work to undertake with respect to system optimization.

The project could be developed and optimized as a more sophisticated system. Respecting the Feature Extraction and Sound Modulation System, an assistant product could be created as an application of this project, which could aid people in navigation and location. Therefore, individuals with vision impairment could benefit a lot from the assistant product. Moreover, it could be developed as an acoustic-assisted recognition system, because in some cases the visual information is difficult to extract in some certain environments.

5 REFERENCES

1. J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series", Math. of Comput., 19, pp. 297-301. (1965).
2. James W. Cooley, Peter A. W. Lewis, and Peter D. Welch, Historical Notes on the Fast Fourier Transform. IEEE Transactions on Audio and Electroacoustics 15 (2), 76-79.(1967)
3. Raghu Muthyalampalli. Implementation of Fast Fourier Transform for Image Processing in DirectX 10. (2011).
4. Cs.unm.edu. Introduction to the Fourier Transform. (2016)
5. MATLAB and Octave Functions for Computer Vision and Image Processing. Peter Kovesi.
6. Ronnie T. Apteker, James A. Fisher, Valentin S. Kisimov, Hanoch Neishlos. Video Acceptability and Frame Rate. IEEE MultiMedia. 2 (3), 32-40. (1995).
7. Modulation. Granite Island Group. (2014).
8. Shrilekha Bangar, Preetam Narkhede, Rajashree Paranjape. Vocal Vision For Visually Impaired. The International Journal Of Engineering And Science. 7. (2013).
9. Eugene P. What Is Image Acquisition in Image Processing. Angela B. wiseGEEK. Conjecture Corporation. 1. (2014).
10. Michelle R. Greene and Aude Oliva. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. Cogn Psychol. (2), 137-176. (2009).
11. Geoff Dougherty. Shape and Texture. Digital Image Processing for Medical Applications. New York: Cambridge University Press. 424-425. (2009).
12. C. Jeanguillaume, C. Colliex. Spectrum-image: The next step in EELS digital acquisition and processing. Ultramicroscopy. 252-257. (1989).
13. Image Entropy - Cassini Lossy Compression Software Tests. Department of Astronomy, Cornell. (2014).
14. Intrinsic Uniformity - Acceptance and reference quality control. Diagnostic Nuclear Medicine. (2014).
15. Michael J. Proulx, Petra Stoerig, Eva Ludowig, Inna Knoll. Seeing 'Where' through the Ears: Effects of Learning by Doing and Long-Term Sensory Deprivation on Localization Based on Image-to-Sound Substitution. (2008).
16. Peter B. L. Meijer. An experimental system for auditory image representations. IEEE Transaction on Biomedical Engineering. 39, 112-121. (1992).