Role of Object Identification in Sonification System for Visually Impaired

R.Nagarajan, Sazali Yaacob and G.Sainarayanan
Artificial Intelligence Application Research Group
School of Engineering and Information Technology
Universiti Malaysia Sabah, 88999 Kota Kinabalu, Malaysia
E-mail: jgksai@webmail.ums.edu.my

Abstract: In this paper the role of object identification in sonification system for Navigation Assistance Visually Impaired (NAVI) is discussed. The developed system includes Single Board Processing System (SBPS), vision sensor mounted headgear and stereo earphones. The vision sensor captures the vision information in front of blind user. The captured image is processed to identify the object in the image. Object identification is achieved by a real time image processing methodology using fuzzy algorithms. The processed image is mapped onto stereo acoustic patterns and transferred to the stereo earphones in the system. Blind individuals were trained with NAVI system and tested for obstacle identification. Suggestions from the blind volunteers regarding pleasantness and discrimination of sound pattern were also incorporated in the prototype. With the object identification, the discrimination of object and background with the sound is found to be easier compared to sound produced from the unprocessed image

1. INTRODUCTION

Most aspects of the dissemination of information to aid navigation and cues for active mobility are passed to human through the most complex sensory system, the vision system. This visual information forms the basis for most navigational tasks and so with impaired vision an individual is at a disadvantage because appropriate information about the environment is not available. The number of visually handicapped persons worldwide would double from the present 45 million by 2020 [1,2].

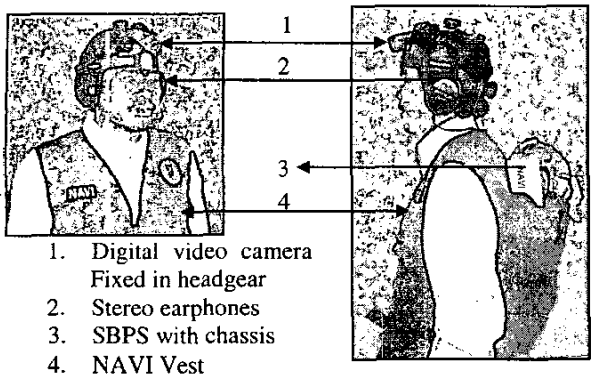
Electronic Travel Aids (ETA) are electronic devices developed to assist the blind for autonomous navigation [10]. The development of vision aid for blinds had been under extensive research from the beginning of 1970's. It has been attempted in many ways with restricted achievement [3,5][1]. Early ETA's use ultrasonic sensors for the obstacle detection and path finding. Recent research efforts are being directed to produce new navigation systems in which digital video cameras are used as vision sensors. Peter Meijer's [6] *The vOICe* is one of the latest patented image sonification system. Video camera is used as vision sensor. A dedicated hardware was constructed for image to sound conversion. The image captured is scanned in the left-right direction with sine wave as sound generator. The top portion of the image is transformed into high frequency tones and the bottom portion into low frequency tones. The brightness of the pixel is transcoded into loudness. Similar works had been carried out by Capelle and Trullemans [4]. All the earlier works in the direction of capturing the image of environment and mapping the image to sound, do not undertake any image processing efforts to provide the information of the objects in the scene. Instead captured image is directly sonified to sound signals. In general, background fills more area in the image frame than the objects, as the sound produced from the unprocessed image will contain more information of the background. It is also noted that most of the background is of light colors and the sound produced on it will be of high amplitude compared to the objects in the scene. This may be the reason for blinds finding difficulties in understanding the sound produced.

In this paper, object identification is achieved using a clustering algorithm. The identified objects are enhanced. Importance is given to the objects in the environment than the background of the environment for sound production. This will enable the blind user to identify the obstacles easier.

2. HARDWARE OF NAVI SYSTEM

The hardware model constructed for this vision substitution system has a headgear mounted with the vision sensor, stereo earphone and Single Board Processing System (SBPS) in a specially designed vest for this application. The user has to wear the vest. The SBPS is placed in a pouch provided at the backside of the vest. SPBS selected for this system is PCM-9550F with Embedded Intel® low power Pentium® MMX 266 MHz processor, 128 MB SDRAM, 2.5" light weight hard disk, two Universal serial bus and RTL 8139 sound device chipset assembled in Micro box PC-300 chassis. The weight of SBPS is 0.7 Kg. Constant 5V and 12 V supply for SBPS is provided from the batteries placed in front packets of vest. Vision sensor selected for this application is a digital video camera, KODAK DVC325. A blind individual carrying the headgear and processing equipment in the vest is shown in Figure 1. The work is progressing to miniaturize the size of the equipment, so as to be more convenient for the blind individual to carry.

[1] 0-7803-7651-X/03/$17.00 © 2003 IEEE



1. Digital video camera Fixed in headgear
2. Stereo earphones
3. SBPS with chassis
4. NAVI Vest

Figure 1. Blind Volunteer with NAVI system

3. OBJECT IDENTIFICATION

Digital video camera mounted in the headgear captures the vision information of scene in front of the blind user and the image is processed in the SBPS in real time. The processed image is mapped to sound patterns. Image processing should be properly designed to have effective sonification. Since the processing is done in real time, the time factor has to be critically considered. The image processing method should be of less computation. In industrial vision system applications, there can be a priori knowledge on the features such as contour or size of the object to be detected; here, with the known features, the object of interest are identified by eliminating the background [8]. But, in the proposed vision substitutive system, the nature of object to be identified is undefined, uncertain and time varying. The classical method for object identification and segmentation cannot be used in this application. The main effort in this module is to identify the objects in the scene in front of the blind. Unless the task is automated, it will be very difficult for the blind user to understand the environment and navigate without collision. One of important features needed by the blind user in the image from the environment are the orientation and size of the object and obstacles. During sonification, that is to be discussed in later section, the amplitude of sound generated from the image directly depends on the pixel intensity. In any gray image, pixel value of white color is of maximum of 255 and black is with minimum of zero. As the image pixels of light color produces sound of higher amplitude than darker pixels. Acoustic pattern set of pixels with bright colors in the dark background is easy to identify than the dark pixels over bright background. It can be felt that the background of the most real world pictures are of bright colors than the object. If the image is transferred to sound without any enhancement, it will be a complex task to understand the sound, which is the major problem faced in early works [4]. There can be a possibility that the background may also have some important features and these features will be eliminated if a total elimination of background is undertaken. Hence an effort is made in this paper to suppress the background instead of elimination and also to enhance the object of interest in order to impart more consideration to the object. Human vision system creates the concentration of vision only on the region of interest, while other regions are considered as background and are given less consideration and focus. Generally the focused object will be in the center of vision and this property of human vision is incorporated in the proposed method. Human vision system creates the concentration of vision only on the region of interest, while other regions are considered as background and are given less consideration and focus. This is the property of iris in the human vision system [11]. There are more chances of object to be in the central (iris) area of vision; if not, the object can be aligned to the central area by moving the iris or by turning the head. In the case of blind, moving the head is appropriate since the central area of vision is fixed to the camera. This is one of the main aspects to be considered during object identification.

3.1 Feature extraction and clustering

The main objective of this work is to suppress the background and to enhance the object; for this, the gray levels of the object and background have to be identified. Image used for processing is of 32x32 pixel size and of four gray levels namely black (BL), white (WH), dark gray (DG) and light gray (LG). Feature extraction is the most critical part in image processing. The extracted features should represent the image with limited data. The type of features extracted from image for object identification or classification depends on the application and also mainly on the computational time. In these work four features are extracted from each gray level. Each image will have four feature vector namely X_{BL}, X_{DG}, X_{LG} , X_{WH} , each with four features such as

$$X_{BL} = [\, x_1, x_2, x_3, x_4 \,]$$
$$X_{DG} = [\, x_1, x_2, x_3, x_4 \,]$$
$$X_{LG} = [\, x_1, x_2, x_3, x_4 \,]$$
$$X_{WH} = [\, x_1, x_2, x_3, x_4 \,]$$

where,

x_1= Represents the number of respective gray pixel in the image, this is a histogram value of the particular pixel.

x_2 = Represents the number of respective gray pixel in the central area of the image. Iris area is the central area of human eye, which maintains a concentration of vision. This concentration is distributed towards the boundary in a nonlinearly decreasing function. The central area of the image obtained by the camera is considered here as iris area and thus models the human eye. Generally the object of interest will be in the center of human vision.

x_3 = Represents the pixel distribution gradient. Its value depends on the location of the particular gray level pixels in the image area. x_3 is calculated by the sum of the gradient values assigned to the pixel location. The gradient value increases towards the center with gaussian function. So that, pixel of a particular gray

level in the center has comparatively higher value than the pixels of same gray level in outer area.

x_4 = Represents the gray value of the pixel. Generally most of the background in the real world are of light colors than the objects.

In this experimentation, 250 data are collected from simulated as well as real life images. The extracted data have to be clustered into two classes namely object class and background class. The number input nodes is 4. The target values for clustering are fixed apriori. Fuzzy Learning Vector Quantization (FLVQ) network is trained with 150 data [7,9]. The trained network is then tested for all 250 data. The data are clustered to the class with minimum euclidean distance with its final weight values. The success of classification has been found to be more than 97.6 %. When the trained network is incorporated in the real time processing, it is able to identify the object and the background. On detection, the object pixels are enhanced, while the background pixels are suppressed with following algorithm

Let G_o be gray level as classified to object class of FLVQ network, G_b be the gray level as classified to background class of FLVQ network and I be the preprocessed image.

$$\begin{aligned} &For\ i, j = 1,2,3,\ldots,32 \\ &If\ I(i,j) == G_o \\ &\qquad then\ I(i,j) = K_1 \\ &If\ I(i,j) == G_b \\ &\qquad then\ I(i,j) = K_2 \\ &End \\ &\qquad I_1 = I \end{aligned} \tag{1}$$

where K_1 and K_2 are chosen scalar constants, $K_1 >> K_2$ and symbol '=='means that RHS and LHS are equal. I_1 be the image with background suppressed and object enhanced.

4. STEREO SOUND GENERATION

The processed image is sonified to stereo acoustic patterns. The sine wave with the designed frequency is multiplied with gray scale of each pixel of a column and summed up to produce the sound pattern. The frequency of the sine wave is inversely related to pixel position and the loudness of the sound is made to depend directly on the pixel value of processed image. The sound pattern from each column of image pixels is appended to construct the sound for whole image. The scanning of picture is performed in such a way that stereo sound is produced.

Let

f_o be the fundamental frequency of the sound generator

G be a constant gain.

F_D be the frequency difference between adjacent pixels in vertical direction.

The changes in frequency corresponding to $(i,j)^{th}$ of the pixel in 32 x 32 image matrix is given by

$$f_i = f_o + F_D \tag{2}$$

$$\text{where } F_D = Gf_o(32 - i);\ i=1,2,3,\ldots,32 \tag{3}$$

In the proposed system, the frequency is linearly varied, by maintaining F_D as a constant.

The generated sound pattern is hence given by

$$S(j) = \sum_{i=1}^{32} I_1(i,j) \sin \omega(i)t\ ;\ j = 1,2,\ldots,32 \tag{4}$$

where,

$S(j)$ is the sound pattern for column j of the image

t = 0 to D; D depends on the total duration of the acoustic information for each column of the image

$\omega(i) = 2\pi f_i$, where f_i is the frequency corresponding to row, i.

In this stereo type scanning, the sound patterns created from the left half side of the image is given to left earphone and sound patterns of right half side to right earphone simultaneously. The scanning is performed from leftmost column towards the center and from right most column towards the center [11].

5. IMPORTANCE OF OBJECT IDENTIFICATION

The importance of the image processing stages undertaken in NAVI can be illustrated by comparing the sound in 3D form for an image with and without image processing. It is important to note that, by the human auditory nature, it is easy to identify and differentiate a high amplitude sound in the middle of low amplitude sound, compared to low amplitude sound in between high amplitude sound [12]. Generally, the background is assumed to take more image area and is of light color compared to object. If the proposed image processing methodology is not undertaken, the background is transformed to high amplitude sound compared to that of object and therefore the features of the background will predominate over the object. The distribution of the frequency and the amplitude in the sound produced from the unprocessed image and processed images are shown in Figure 2 and Figure 3. Image considered is split into left half and right half images namely I_L and I_R respectively. Distribution of sound S_L to the left ear phone and S_R to the right earphone are shown in three dimensional plot (3D), in which x axis represents the time after the starting of sound, y axis represents the frequency and z axis represents the amplitude. In Figure 2, the sound from the background predominates the sound produced from the objects. This may cause confusion for the blind user to discriminate the object from the background. In Figure 3,

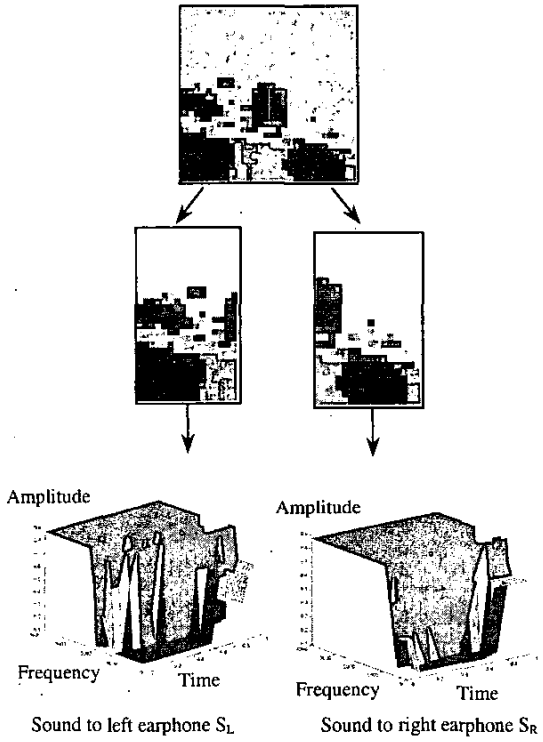


Figure 2. 3D plot of sound from unprocessed image

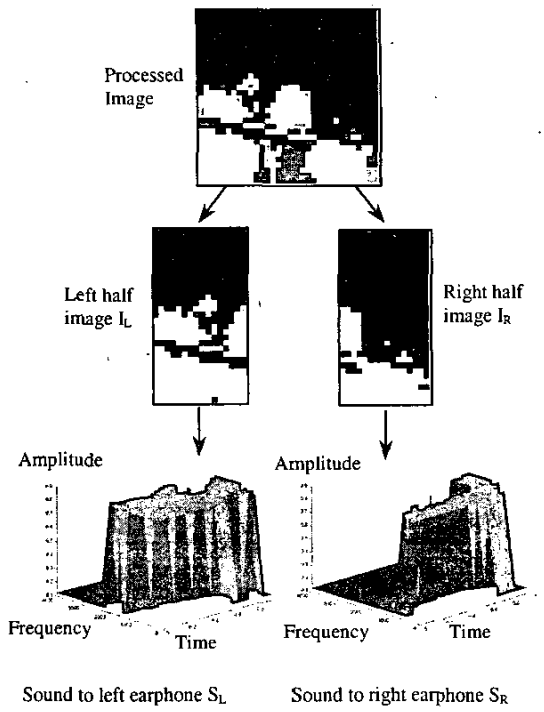


Figure 3. 3D plot of sound from processed image

the sound from the objects predominates over the sound from the background, as it will be easier for discrimination. From the above examples and discussions, the importance and necessity of the proposed object identification module can be acknowledged.

6. CONCLUSION

The developed prototype headgear and scheme was tested on blind persons. The blind persons were trained with some basic geometric shapes and to identify obstacles of in-door environment. The image processing designed for NAVI is found to be suitable for this application. With the proposed object identification method, blinds were able to identify the objects in the environment with less effort. It was encouraging to note that, he is also able to find the objects moving with a nominal speed. Work is continued to train the blind person in identifying the outdoor scene through the sound pattern produced by this prototype. In this research, information regarding depth of the object is not considered. However by comparing the sound patterns from relative distances between the blind person and the object, information regarding the nearness of objects can be manipulated by the blind person after getting an experience with the developed scheme. That is, an object is 'perceived' bigger through the variation in sound pattern as the blind moves near to the object.

Acknowledgment: Authors wish to thank MOSTE for funding the research through Universiti Malaysia Sabah: IRPA code: 03-02-10-0004

REFERENCES

[1] World Health Organization (WHO), 1997, Blindness and Visual Disability, Part I of VII: General Information , *Fact Sheet 142.*

[2] ERM : Ethnologue Report for Malaysia. 2001. http://www.ethnologue.com/show_country.asp?name =Malaysia+%28Peninsular%29

[3] Farrah Wong, R.Nagarajan, Sazali Yaacob, Ali Chekima and Nour Eddine , "Electronic Travel Aids for Visually Impaired – A Guided Tour", *Conference in Engineering in Sarawak*, Proceedings pp 377-382, 19-20, May 2000, Malaysia

[4] Christian Capelle and Charles Trullemans, "A Real-Time Experimental Prototype for Enhancement of Vision rehabilitation Using Auditory Substitution", *IEEE Trans. on Biomedical Engineering*, Vol 45, No. 10, pp 1279-1293, Oct 1998.

[5] Fish, R. M., "Auditory display for the blind", US Patent No. 3800082, 1974.

[6] Peter B.L. Meijer," An Experimental System for Auditory Image Representations", *IEEE Transactions on Biomedical Engineering*, Vol 39, No. 2, pp 112-121, Feb 1991.

[7] James C.Bezedek, James Keller, Raghu Krishnapuram and Nikhil R.Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, Boston, 1999.

[8] Nikil.R.Pal and Sanker.K.Pal, A Review on Image Segmentation Techniques, *Pattern recognition*, Vol. 26, pp 1277-1293, 1993.

[9] L.Faussent, *Fundamentals of Neural Networks*, Prentice Hall, New Jersecy, 1994.

[10] National Federation of the Blind (NFB). 2002. http://www.nfb.org/default.htm

[11] G.Sainarayanan, R.Nagarajan and Sazali Yaacob, "Incorporating Certain Human Vision Properties in Vision Substitution by Stereo Acoustic Transform" Proceedings of IEEE Sixth International Symposium on Signal Processing, ISSPA 2001. 13- 16 August 2001, Malaysia.

[12] Perrott, "Discrimination of the spatial distribution of concurrently active sound sources: Some experiments with stereophonic arrays". *Journal of the Acoustic Society of America*, 76(6), 1704-1712, 1984.