Object Discrimination using Stereo Vision for Blind through Stereo Sonification

G.Balakrishnan G.Sainarayanan R. Nagarajan Sazali Yaacob
AI Research Group, School of Engineering and Information Technology
Universiti Malaysia Sabah, 88999, Kotakinabalu, Malaysia
Tel: +6088-320000 x 3048, Fax: +6088-320348
E-mail: g_krishbalu@yahoo.com

Abstract

In this paper, a development of image processing, stereovision methodology and a sonification procedure for image sonification system for vision substitution are presented. The hardware part consists of a sunglass fitted with two mini cameras, laptop computer and stereo earphones. The image of the scene in front of blind people is captured by stereo cameras. The captured image is processed to enhance the important features in the scene in front of blind user. The image processing is designed to extract the objects from the image and the stereo vision method is applied to calculate the disparity which is required to determine the distance between the blind user and the objects. The processed image is mapped on to stereo sound for the blind's understanding of the scene in front. Experimentations were conducted in the indoor environment and the proposed methodology is found to be effective for object identification, and thus the sound produced will assists the visually impaired for their collision free navigation.

1. Introduction

Navigation is one of the most important aspects of human life to meet many goals in daily life. Information to aid navigation is through the human vision system. With impaired vision an individual is at a disadvantage, because appropriate information about the environment is not available. An autonomous collision free navigation with discrimination of objects becomes major problem faced by the vision impaired. According to World Health Organization census, around 180 million people worldwide are visually disabled; of those 40 to 45 million populations are totally blind [1]. This population is expected to double by the year 2020.

Electronic Travel Aids (ETA) are equipments that help to present visual information to visually impaired people, so that a visually impaired person can have a closer interaction with the environment. Many devices exist to assist visually impaired people for navigation. The early version of ETA devices [2] utilizes ultrasonic sound or laser and provides tactile or auditory feedback to the blind about their immediate environment. These devices often provide little benefit in mobility because of improper calibration of sensors, difficulty in holding the ETA and complexity in understanding their feedback signals. Recent research efforts are being directed to produce new navigation systems in which digital video cameras are used as vision sensors.

Peter Meijer's *The vOICe* [3] is one of the latest patented image sonification system. It has introduced the concept of using video camera instead of ultrasonic or laser sensors. The camera is mounted on to a headgear of the user. The image captured by the camera is resized and gray scaled. The gray scaled image is sonified with sine wave function. The scanning of image is done from left to right for sonification. 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 converted into loudness. The sound generated is transmitted to earphones of the user and thus the environment is perceived. A dedicated hardware was constructed for image to sound conversion. Similar works had been carried in Navigational Assistance for Visually Impaired (NAVI) [4].

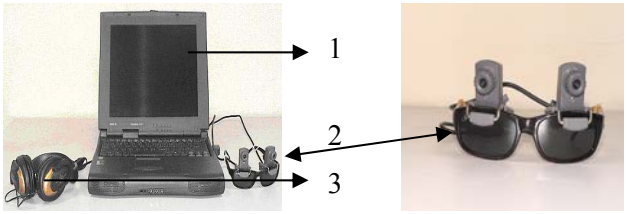
In NAVI, the scene infront of the user is captured using the head mounted camera and the captured image is resized to 32 X 32 and the gray scale is reduced to 4 levels. The image is differentiated into objects and background. The objects are assigned high intensity of gray value and the background is assigned with low intensity. The processed image is converted into stereo sound, where amplitude of sound is directly proportional to intensity of image pixels, and frequency of sound is inversely proportional to vertical orientation of pixels. Both in *The vOICe* and NAVI the distance of the obstacle from the user cannot be obtained directly. The distance is one of the important aspects for collision free navigation for blinds. With a single camera, image intensities do not typically correspond to depth information, which is necessary for navigation. In order to incorporate the distance information, stereo cameras are used in this work.

Use of stereo vision sensors has been of research interest mainly for extraction of 3D models of objects and for perception of depth. The manner in which human

beings use their two eyes to see and perceive the three-dimensional world has inspired the use of two cameras to model the world in three dimensions. The different perspectives of the same view seen by two cameras (displaced by a known distance) lead to a relative displacement of the same objects or the same points in world reference (called disparity). The size and direction of these disparities can be utilized for depth estimation. The depth of a point is inversely proportional to the amount of disparity.

2. Experimental Prototype

A prototype system is designed for this work. The hardware model constructed for this vision substitution system has a sunglass fitted with two mini video cameras, stereo earphone and a laptop computer. The two cameras are displaced from each other by a distance of 5 cm. Both the cameras are adjusted to same focus by experimentations. These cameras are connected to laptop through USB ports. Figure 1 (a) show the experimental setup used in this work and Figure 1 (b) shows the stereo camera on a sunglass. When the sunglass is weared by the blind user, the stereo cameras capture the scene infront of the user.



(a) (b)

1 – Laptop Computer
2 – Sunglass fitted with stereo camera
3 – Stereo headphone

Figure 1: Prototype system

3. Image processing methodology

The blind users will be confused if the entire environmental information is provided to them with equal preference [5]. Information has to be optimized so that only the essential environmental information is made available to the user. Image processing is performed to satisfy the above constraints. The proposed image processing methodology identifies the objects from both left and right images to form images with only objects. The steps involved in isolating objects from the images are explained in the following sections.

3.1. Initial processing

The first step in this methodology is image acquisition, in which scene infront of blind is captured using both the cameras simultaneously. The image captured from camera is color image of size 352 x 288. Processing the image with original size will increase computation time. Blind navigation requires real time processing. The time factor is very critical in this application. The image processing method has to be done with less computation. Therefore pre-processing is undertaken to reduce the computation time, where both the left and right stereo images are converted to gray scale intensity image and resized to 64 x 64. Figure 2 shows initial processed left and right image of real life picture with three objects.

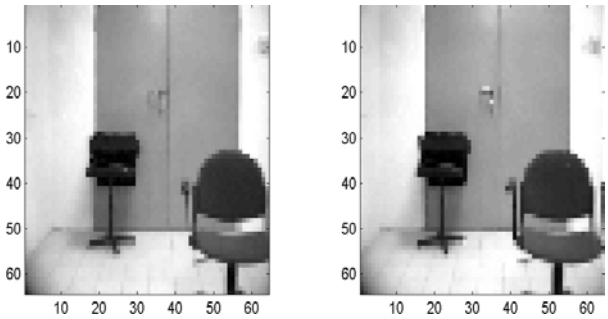


Figure 2: Left and right processed image

The main task is to identify and assign importance to the objects based on its distance. In this work, the objects in both the images are identified by locating its edges.

3.2. Boundary extraction

Edge detection is one of the important human vision properties as it has the ability to recognize the object boundaries. The goal of edge extraction is to provide useful structural information about object boundaries. Canny edge detector is used for edge detection because this method uses two thresholds to detect strong and weak edges, and includes the weak edges in the output only if they are connected to strong edges. Canny edge detector is an optimum edge detector [6]. Figure 3 shows the edge features of the right camera image extracted using canny edge detector. Similar processing is undertaken for left camera image also.

Figure 3: Edge features of right camera image

3.3. Broken Boundary Linking

In this work, the region within the closed boundary is considered as an object. After edge detection, it is found that the edge features are incomplete and discontinuous due to resizing of image and lighting effects. So it is realized that edge detection alone is not enough to extract the closed boundary of an object from the image. Therefore further processing is undertaken to connect the edges into meaningful object. Edge linking process is required to assemble these edges into continuous and closed edges.

Morphological operation using dilation is proposed to link the broken boundaries. By experimentation it is found that edges of the same objects are broken by a maximum of 2 pixels. So a horizontal and vertical disk-structuring element with the size of two pixels is created. This structuring element is used in dilation operation to connect the broken edges around the boundary of all objects. Thus the edges around the object that are broken by 2 pixels in horizontal and vertical directions are connected. Figure 4 shows the result of edge linking on Figure 3.

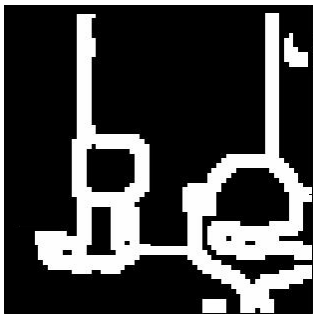


Figure 4: Image after edge linking

3.4. Noise removal

After edge linking, the boundaries of objects will be closed and made continuous. The intensity of pixels inside the closed boundary is enhanced to higher level using flood fill operation. Each object in the image is labeled. Some of the edges other than objects still exist. These edges have to be removed. The goal of noise removal is also to remove extraneous edges present in the image without eliminating the desired objects. Morphological operations using a combination of erosion and dilation operations are undertaken to remove the superfluous edges.

The size of disk-structuring element is selected in such a way that it removes only the noise when eroding the image. By experimentation a structuring element with the size of three pixels is created. This structuring element is used in erosion operation to remove three pixels from around the boundary of all objects. As a result, extraneous edges presents in the image will be eliminated and the objects will be shrinking. To restore the objects to their original size, dilation operation is applied to the eroded objects using the same structuring element. Thus binary image with only objects will be obtained. Figure 5 shows the binary image with only objects of right camera image.

Figure 5: Binary image after noise removal

4. Isolated Object Image

Object identification steps are applied on both left and right image to obtain binary image with only objects. With binary image stereo matching will not be effective. These binary images are mapped with the resized gray scale intensity images and new gray scale intensity image with only objects of both left and right camera are extracted. Isolated gray scale intensity object images of left and right camera is shown in Figure 6. Isolated object image reduces the computation time of stereo matching as only the objects are considered for matching.

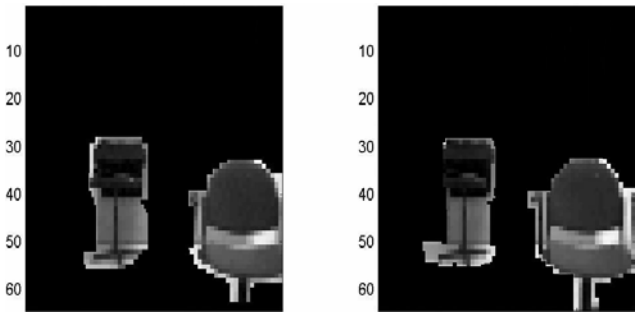


Figure 6: Left and right intensity image with only objects

5. Stereo vision methodology

In order to calculate the distance of the object, the disparity has to be computed using stereo vision concept. The basic idea of stereo vision [7] is illustrated in Figure 7, Assume that a point P in a surface is projected on two cameras image planes, P_L (x_l,y_l) and P_R (x_r,y_r)

respectively. Let O_L is the optical center of left camera, O_R is optical center of right camera. Here both camera coordinates axes are aligned, and the baseline (line segment joining the optical centers of two cameras) is parallel to the camera x coordinate axis.

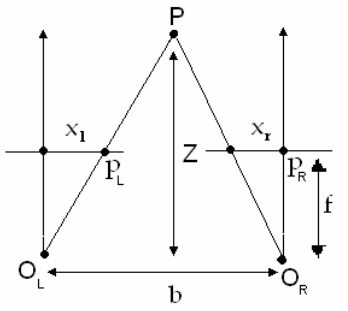


Figure 7: Stereo Geometry

Given the baseline T (distance between O_L and O_R), and the focal length f of the cameras, depth at a given point may be computed by similar triangles as

$$Z = f\frac{b}{d} \quad (1)$$

where z is the depth of point P and d is the disparity of that point, $d=x_l - x_r$ [8].

Given the left and right stereo object images, area based stereo matching is performed. Since the cameras are adjusted to same focus by experimentation, the search for correspondence is reduced to a 1-D search. Stereo correspondence is established using SSD (Sum of Squared Differences) correlation method [9]. For each left image pixel, its correlation with a right image pixel is determined by using a small correlation window of size 5*5 pixels in which we compute the SSD of pixel intensities as

$$c(\vec{d}) = \sum_{k=-W}^{W}\sum_{l=-W}^{W}\psi(I_l(i+k,j+l)I_r(i+k-d_1,j+l-d_2)) \quad (2)$$

where *(2W+1)* is the width of the correlation window. I_l and I_r are the intensities of the left and right image pixels respectively. [i, j] are the coordinates of the left image pixel. $\vec{d} = [d_1, d_2]$ is the relative displacement between the left and right image pixels.

$\psi(u,v) = -(u-v)^2$ is the SSD correlation function.

Here, the search is reduced to 1-D search. Therefore, d_2 is always equal to 0. The correlation-matching algorithm consists of calculating the correlation values pixel by pixel by varying d_1, which slides the correlation window from the left to the right in the right image along this row. The corresponding point is determined by the pixel that has the highest correlation value. The disparity between the two corresponding points is the distance d_1 that separates them. Since only the isolated object images are used for correspondence, the mismatch error is less and also the computation time is reduced when compared to conventional area based and feature based techniques.

The disparity value of some pixels within the same object varies due to mismatch. To have a uniform disparity value for an object, histogram is used, where the disparity value that occurs most within the object is found and that value is assigned to all pixels within the object. It is obvious that objects close to the user have high disparity compared to far objects. Intensity values are assigned to objects based on their disparity. The objects having high disparity value are assigned with high intensity of gray value and the objects with low disparity value are assigned with low intensity value. Figure 8 shows the final disparity map after uniform assignment.

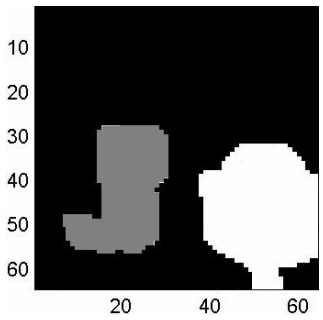
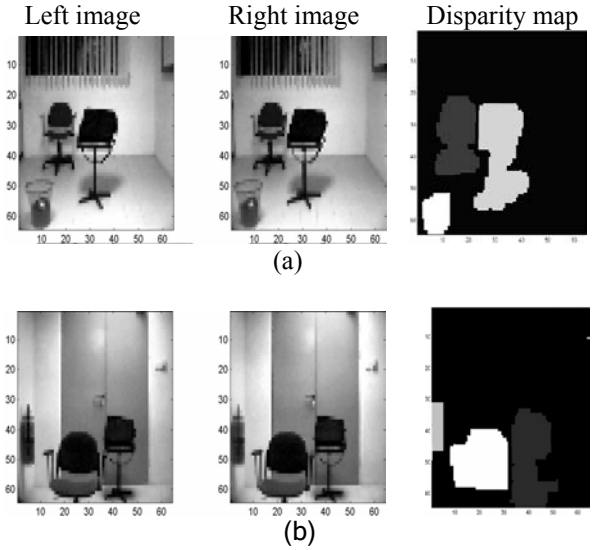
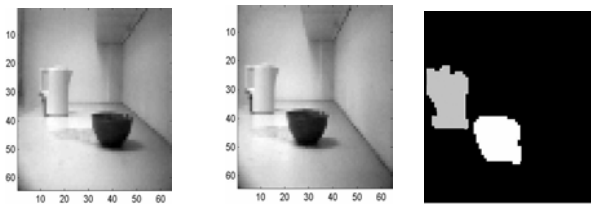


Figure 8: Final disparity map

Following Figure 9 shows some of the results obtained from the stereo pairs taken in the indoor environment. Inspite of different backgrounds, the proposed scheme identifies the objects effectively.





(c)

Figure 9: Results of some indoor stereo pairs

The disparity map obtained is converted into stereo sound using the sonification method. Transformation of data in relation to perceived associations to an acoustic signal for the purpose of facilitating communication or interpretation is defined as Sonification [10]. Human auditory perception is, in particular, sensitive to changes in sound, as it is well suited to discriminate between periodic and non-periodic events and to even detect small changes in the frequency of continuous signals. Although there are only a small number of sound characteristics, they can be manipulated to produce a rich set of sounds. One can change the amplitude, frequency, or shape of a sound wave, and each of these variations reflects psycho acoustical effects [11].

Human auditory system can sense frequencies between 20 Hz to 20 KHz [12], but from literature and experimentations it is observed that the system is most sensitive to frequencies between 20 Hz to 4000 Hz. This range is adopted in the proposed sonification module. The disparity image size is resized to 32 x 32 so that the computation time in sound generation is reduced and made approximate to the computation time of the proposed methodologies. Let the new resized image be I_1. In order to create variations in pitch in the sonification module, the pixel position in a column of the image pattern is made to be inversely related to frequency of sine wave. In this work, the loudness is made to depend directly on the pixel value of processed image and thus the pixel position of the image data is sonified through the stereo earphones.

Here, the image data is sonified to stereo sound by proper mapping of the image, such that the information regarding image data corresponding to left side of a blind are transferred to the left earphone and the right half image data are transferred to the right earphone. Let f_o be the fundamental frequency of the sound generator, G be a constant gain and 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

$$f_i = f_o + F_D \quad (3)$$

where $F_D = Gf_o(32 - i)$; i=1,2,3,…,32 (4)

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

The generated sound pattern is hence given by

$$S(j) = \sum_{i=1}^{32} I_1(i, j) \sin \omega(i)t \text{ ; j = 1,2,…,32} \quad (5)$$

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.

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 sound pattern from each column is appended to construct the sound for whole image. The scanning of picture is performed in such a way that stereo sound is produced. 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.

Sound pattern to the left earphone is S_L= S(1) to S(n/2) appended from the left side. Sound pattern to the right earphone is S_R= S(n) to S(n/2) appended from the right side, where n = 32, is total number of columns.

6. Conclusion

In this paper, a scheme for obstacle detection and auditory transformation applied in navigation aid for visually impaired are discussed. This work makes use of image processing module to differentiate the objects or obstacles from the image and stereo vision module to classify the obstacles based on its distance from the user. Different intensity of gray value is assigned to the objects based on its disparity and the image is converted into sound using sonification procedure. Since only the object image is used for stereo matching, mismatch error is reduced and the computation time is minimized. Experiments were conducted in indoor environments and the sound pattern generated provides some valuable information about the environment. In future the work is extended to make the system more portable and training the blind people.

Acknowledgment

Authors wish to thank Ministry of Science, Technology and Environment, Malaysia for funding the research through Universiti Malaysia Sabah under IRPA code: 03-02-10-0043/EA0041.

7. References

[1] "*World Sight Day: 10 October*", World Health Organization, 10 October 2002, URL: http://www.who.int/mediacentre/releases/pr79/en/print.html

[2] Wong, F., Nagarajan, R., Yaacob, S., Chekima, A. and Belkhamza N-E, *Electronic travel aids for visually impaired – A guided tour,* Proceedings of Conference in Engineering in Sarawak, Malaysia, pp 377-382, 2000.

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

[4] Sainarayanan, G., *On Intelligent Image Processing Methodologies Applied to Navigation Assistance for Visually Impaired*, Ph. D. Thesis, University Malaysia Sabah, 2002.

[5] R.Nagarajan, Sazali Yaacob and G.Sainarayanan, *Role of Object Identification in Sonification System for Visually Impaired*, IEEE TENCON 2003-International Conference, October 2003.

[6] Canny, J., *A Computational Approach to Edge Detection,* IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 6, pp. 721-741, Nov 1984.

[7] Myron Z.Brown and Gregory D.Hager, *Advances in computational stereo*, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 25 No.8, Aug 2003, pg 993-1008.

[8] David A.Forsyth and Jean Ponce*, Computer Vision: A modern approach*, 2nd Ed. New Jersey: Prentice-Hall, 2002

[9] Umesh R.Dhond and J.K.Aggarwal, *Structure from stereo*, IEEE Trans. On systems, man and cybernatics, Vol. 19, No. 6, Nov/Dec 1989.

[10]Altman, J. A., *Role of the higher parts of the auditory system in the location of a moving sound source,* Journal of Neuroscience and Behavioral Psychology, 14(3), 200-205, 1984.

[11] Frysinger, S. P,, *Applied research in auditory data representation*, Proceedings of the SPIE, Conference 1259, Extracting Meaning from Complex Data: Processing, Display, Interaction, pp. 130 – 139, 1990.

[12] ICAD (Int. Community for Auditory Display), "*Sonification report: status of the field and research agenda prepared for the National Science Foundation*", http://www.santafe.edu/~icad/, 1997.