31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

SoundView: An Auditory Guidance System Based on Environment Understanding for the Visually Impaired People

Min Nie, Jie Ren, Zhengjun Li, Jinhai Niu, Yihong Qiu, *Member, IEEE,* Yisheng Zhu, *Senior Member, IEEE,* and Shanbao Tong, *Member, IEEE*

***Abstract*—Without visual information, the blind people live in various hardships with shopping, reading, finding objects and etc. Therefore, we developed a portable auditory guide system, called SoundView, for visually impaired people. This prototype system consists of a mini-CCD camera, a digital signal processing unit and an earphone, working with built-in customizable auditory coding algorithms. Employing environment understanding techniques, SoundView processes the images from a camera and detects objects tagged with barcodes. The recognized objects in the environment are then encoded into stereo speech signals for the blind though an earphone. The user would be able to recognize the type, motion state and location of the interested objects with the help of SoundView. Compared with other visual assistant techniques, SoundView is object-oriented and has the advantages of cheap cost, smaller size, light weight, low power consumption and easy customization.**

I. Introduction

Some retinal diseases could permanently result in the complete loss of vision, which is impossible to recover with current medical techniques and therapies. Therefore, developing approaches for partly restoring the vision or helping the blind to understand the environment has been an interesting topic for biomedical engineers. However, most artificial vision techniques resort to the stimulation of the visual pathway using epi-retinal, sub-retinal, optic never or cortical microelectrode array, which are all invasive. These visual prosthetics, which is still far from clinical applications, require expensive hardwares under high risk of neurosurgery, and furthermore, they can not be used for the patients who lost their vision in the early ages [1]. On the other hand, blind-assistant facilities have been essentially important for those with impaired vision. Traditionally, the blinds use white canes or guide dogs to explore the environments. However, white cane could provide very little information within a limited distance, while guide dogs are not affordable for most blind patients [2]. In 1970s, the blind-assistant facility based on sensory substitution was introduced [3]. In such a noninvasive rehabilitation method, visual information is transformed into an intact sensory modality [4]. In 1990s, the idea of auditory substitution of vision was put forward. This technique doesn't need surgical operation, and moreover, the system is easy for upgrade and customization. There have been devices based on the "pixel-to-soundel" strategy to generate the sound that represents the objects in environment, which translates, for each pixel, the position into frequency, and brightness into oscillation amplitude [5].

Such a "pixel-to-soundel" technique has the shortcoming of inability to distinguish a single object from the background because each pixel only contains primary physical information and the auditory signal is simply the summation of one column of "soundels". The user has to translate the sound back into the actual objects by him/herself, which means the user must be trained to "decode" the synthesized sounds [6]. In addition, "pixel-to-soundel" may distract the user and influence the normal function of the auditory system.

Other devices, like NavBelt by J.Borenstein [7] which detects obstacles with ultrasonic and give a propositional safe direction to the user with acoustic stereo signals, only provide simple navigation assistant to the blinds while walking. Kawai and colleagues developed a system to detect moving objects and inform the user with stereo sound, but it would be helpful just in specific conditions like ball games [8].

In this paper, we proposed an environment understanding approach based on sensory modality substitution for blind guidance and developed a portable prototype system, i.e. SoundView.

II. System & Architecture

A. Principle of the SoundView System

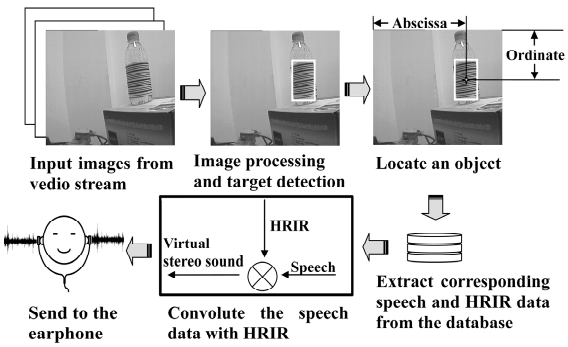

Fig. 1. Principle of A-Sight.

We proposed a new modality conversion approach based on digital signal processor (DSP), which detects the targets in the environment by pattern recognition of image, and then

This work is partly supported by Texas Instruments Inovation Fund.

M. Nie, J. Ren, Z. Li, J. Niu, Y. Qiu and Y. Zhu are with the Department of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, P.R. China

S. Tong is with the Med-X Research Institute, Shanghai Jiao Tong University, Shanghai 200030, P.R. China (stong@sjtu.edu.cn)

encodes the recognized objects into stereo auditory signals for the blind. The schematic system is shown in Fig.1. This project is accordingly called Auditory Sight (A-Sight). By imaging processing, the system obtains the identity, location and motion information of one or multiple objects detected, so that the blind may understand the environment through the auditory modality. Compared with other assistive devices for the blind, the output of A-Sight is more comfortable and more natural.

B. *Architecture Overview*

A prototype A-Sight system, called SoundView, has been developed (Fig.2), which consists of a CCD camera (3.5-8.0mm, 1:1.4, 1/3'' CS, AVENIR CCTV LENS, Nanjing, P. R. China), a DSP platform (ICETECK-DM642-C, Realtime DSP, Beijing, P. R. China) and an earphone. To avoid the complicated pattern recognition, all interested objects in the environment are tagged with barcodes linked with the detail information of the objects in an embedded database. Visual information of the environment is acquired by the CCD camera fixed on a glass through video stream, which is then transferred to the DSP platform. The embedded software detects the barcodes attached on the objects in the images and generate stereo audio signal. The user could thus recognize and locate the object in the environment and its motion status as well.

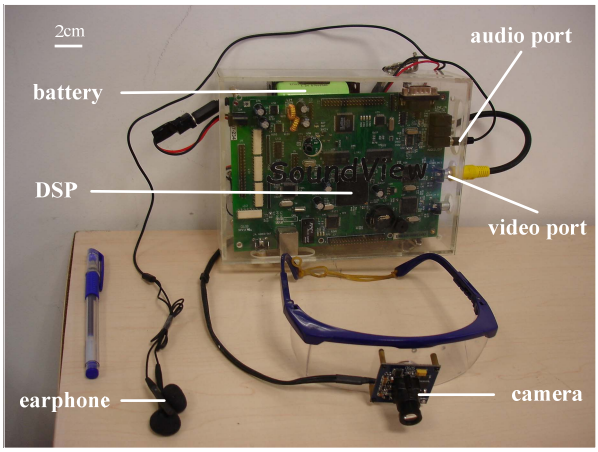

Fig. 2. A prototype SoundView system. The CCD camera connected to the video port of the DSP platform is fixed on a glass to be worn by the user. The earphone is connected to the audio output port of the platform. A battery package is also included. The blue pen on the left side is the scale.

SoundView could be used indoors or outdoors. The whole system is weighted ~2lbs including a built-in battery package. The software in the DSP platform is composed of barcode recognition and auditory encoding. In the recognition stage, barcodes are detected using image processing methods, and the corresponding location information is also saved. In the next auditory encoding stage, SoundView extracts the speech data corresponding to the barcode from an embedded data base and generates stereo auditory signal with head-related transfer function (HRTF) before sending it to the earphone.

C. *Hardware Introduction*

The central unit of SoundView is a TMS320DM642 DSP (Texas Instruments, Dallas, Texas). This system (20cm by 15cm) consists of a 4M*64-bit synchronous dynamic random access memory (SDRAM), a 32M (bit) flash memory, an audio en/decoder chip and SAA7115/SAA7121 video de/encoder chips. Three standard video ports and three audio ports are available on TMS320DM642 as input and output digital media interfaces. Multi-channel audio serial port (McASP) could be extended with software and audio en/decode chips.

SoundView initializes itself when the DSP receives the reset signal or at startup. The algorithms and the database are burnt into flash memory to be executed in the SDRAM. Once reset, the whole system clears all the registers and memory, then the algorithm is reloaded from the flash memory automatically.

Two rechargeable battery packs (5V and 12V) are included for the DSP and the CCD camera respectively. After fully charged, the batteries could support the system for more than one hour which is enough because the user doesn't need to keep the system working all the time.

The software is programmed using Code Compose Studio on PC. The object machine code from the compiler could be emulated on-line and then burnt into the embedded flash memory through the JTAG port, which is easy to debug and upgrade.

D. *Object Detection Based on Barcode Recognition*

Each interested object should be tagged with a unique barcode. In our experiments, the barcodes were in form of interleaved 2 of 5 for up to 100 different objects. Such a barcoding scheme reduces errors in image processing because each code has a start and a terminator sequence.

After analog to digital conversion (ADC), the digital video stream (BT656) from CCD camera is transferred to a temporary buffer. The software analyzes each frame (720 pixels by 576 pixels) as below:

1) Texture detection

(i) Differentiating the image in the buffer, since barcodes are in specific orientations, therefore, the raw image I is differentiated by rows in order to detect the barcodes first.

(ii) Removing the isolating noise points by smoothing the differentiated image with sliding window.

(iii) Since the grey level of the barcodes in the image is influenced by its distance, we implemented two thresholding strategies to detect both close (<4m) and far (>4m) objects:

S1: For those close objects, the differentiated image is binarized with the average grey level (I_{th1}) of the barcode areas just detected as the threshold. The initial threshold is randomly selected (I_{th1}=120 in this paper). After the binarizatoin, those areas {T(i)} with high density of "black points" are detected for the following barcode reading;

S2: While the far objects could be located if there are sharp changes in the images from CCD camera, which can be simply detected by binarizing the differentiated image with a threshold (I_{th2}) dependent on the illumination, e.g. I_{th2}=25 in our experiments.

2) Barcode reading

(i) We then segment the original image *I* to get the areas {A(i)} corresponding to {T(i)}.

(ii) Binarizing {A(i)} into {B(i)} using the average grey level (a_i) of {A(i)} as the threshold. a_i in the current frame will also be saved to update the I_{th1} in *S1*.

(iii) Removing those isolated noise points in {B(i)} with sliding window smoothing as above.

(iv) After above procedures, the start and terminator sequences for the corresponding barcode can be easily detected from smoothed B(i), otherwise, the next area B(i+1) will be analyzed till a barcode is detected.

E. Virtual Stereo Sound Interface

The output of SoundView is stereo auditory representations of the detected objects and their location information, which is simple and user-friendly compared with the traditional "pixel-to-soundel" coding scheme. The user needn't extra time to decode the sound. However, the linguistic voice takes relatively longer time (about 800ms for a word) and the user has to wait until the sound finishes.

Like most of auditory substitution systems nowadays, SoundView also uses head-related transform function (HRTF) to create 3D sound for indicating the location of the object. HRTF combines the interaural time difference (ITD), interaraul level difference (ILD), and the pinna effects [9] for locating the sources in a 3D space. The HRTF database in SoundView is from Center for Image Processing and Integrated Computing (CIPIC) Interface Laboratory in UC Davis [10]. In time domain, HRTF is represented in form of head related impulse response (HRIR). Suppose ***x(n)*** is the speech signal, e.g. the name of the detected objects, then the 3D auditory signal can be generated by convoluting HRIR with x(n):

$$Y_l(n) = \sum_{k=0}^{199} HRIR_l(k) * x(n-k) \qquad (1)$$

$$Y_R(n) = \sum_{k=0}^{199} HRIR_R(k) * x(n-k) \qquad (2)$$

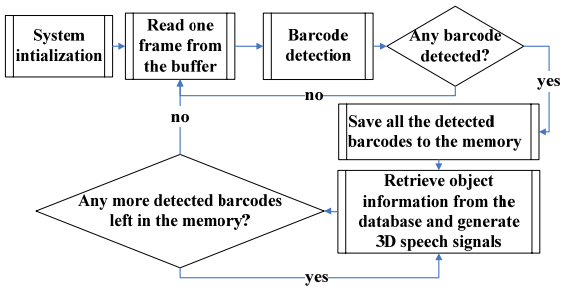

Fig. 3. The flowchart of embedded software in SoundView.

where $Y_l(n)$ and $Y_R(n)$ are the sound signal output to the left and right channel of the earphone. $HRIR_l(k)$ and $HRIR_R(k)$ are the ***k***th samples of HRIR in the left and right channel respectively. Both HRIR and speech signal have the sample rate of 44.1 kHz. For each detected barcode, SoundView would generate a pair of $Y_l(n)$ and $Y_R(n)$ [10].

In CIPIC database, HRIRs in both channels are sampled at 1250 equally-intervaled positions in a spherical surface (radius=1 m). For a specific location, the system would find the nearest sample position in CIPIC database and use the corresponding HRIR data.

Upon no barcode is detected, SoundView will skip the current frame and analyze the next one in the video stream. The flowchart of the image processing and pattern recognition in SoundView is demonstrated in Fig. 3.

As all the barcodes tagged on the objects are in standard sizes, we are able to roughly estimate the distance between the target and the camera according to the scale of barcode in the image captured. The distance of the object therefore can be represented with the amplitude of the sound.

III. Experimental Results

A. Barcode Detection Results

To evaluate the performance of SoundView, laboratory experiments were conducted under uniform illumination by a white heliolamp in an indoor environment. Ten barcode-tagged targets were placed on a table with other objects around. In this test, we used long focal length lens. To get best imaging quality, we kept adjusting the focal length finely during the test. The barcode we used in this test was 9 cm by 10cm. The accuracy for barcode detection at different distances between the camera and object was tested. As shown in Fig.4, barcodes could be successfully detected in a distance range of 50-650 cm. In Fig. 4, it can also be seen that the detected area is lower at a medium distance around 420 cm.

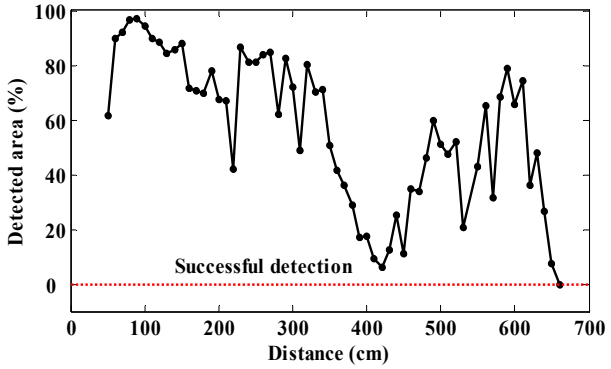

Fig. 4. Experimental results. This figure shows the relationship between the percentage of detected area in a barcode and the distance from camera to the barcode. As long as the rate of detected area is above 0, the barcode could be successfully detected. From this figure, we can see that barcodes could be detected in the full range of 50-650 cm.

B. Time Cost of Image Processing for Each Frame

The image process on DSP is very complicated. Although the parallel processing and complier optimization are implemented, real time processing is one of the main challenges in practice. As shown Fig.5, the processing time for each frame is about 74 ms, i.e. more than 13 frames could be processed per second. The user may not move as fast as the normal people under such a speed, but it is enough for essential indoor activities.

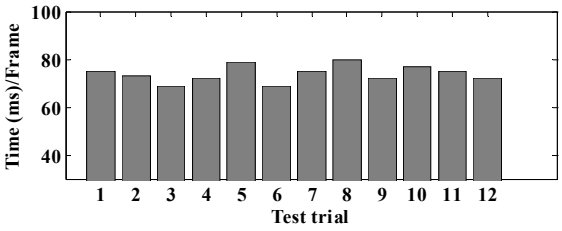

Fig. 5. Average time for processing a frame. The time is estimated from 1000 frames in each test.

IV. Discussions

The barcode detection based object recognition provides fast and accurate processing results, which is not likely affected by the circumrotation of barcodes in vertical plane. For example, if the barcodes are partly covered by other objects, they still could be recognized as long as a line crossing all bars could be fully captured by the CCD camera. Also, adjacently placed objects could be easily differentiated from each other because they are tagged with different barcodes.

In one of the tests, we found that the detected area is low when the distance between the barcode and camera is 420 cm. It was because we used two different strategies to detect the texture feature for close and far barcodes. For those barcodes in medium distance (e.g. d=4.2 m for the camera in this study), neither strategies could perform perfectly. Also, we kept adjusting the focal length to get best image quality, which might be inconvenient for practical application. In the future, this could be avoided by using auto-focus camera. But still, this system can just detect objects in the vision of the camera.

Virtual stereo speech with HRTF and volume modulated distance could not only inform the user what the detected objects are, but also indicate the position information. In indoor environments, our result of barcode detection showed a satisfactory processing speed and sufficient detection range. Such a system may be useful in home life or in a supermarket environment where the objects can be tagged with barcodes. However, barcode is an infeasible way as the identification in real outdoor environment. Therefore, new techniques based on computer vision and automatic pattern recognition should be implemented

In this present prototype system, we use the scale of barcodes in the image to roughly estimate the distance between the targets and the camera, which of course saves a lot of computing load. But for more accurate distance estimation, two or more cameras might be introduced for binocular distance measurement [8].

In addition, our prototype SoundView system was developed on a DSP platform for general purpose. Such a general DSP platform is convenient and efficacious for system developing with abundant hardware resources, e.g. serial ports, ether net ports, and extra video/audio ports, which can be removed to have a much smaller and portable system with lower power consumption.

V. Conclusion

We have presented a prototype auditory guidance system, i.e. SoundView, for the visually impaired people based on the idea of environment understanding. The whole system is based on a portable DSP platform with built-in power supply. Using image processing techniques for barcode detection and HRTF, SoundView provides the identity, location and motion information of the interested objects for the blind user through virtual stereo speech. With the advantages of portability, cheap cost and easy use, SoundView may help the blind people in home, supermarket or the environments when the objects are known. We also tested the usability of the system.

References

[1] A. Jacomuzzi, and N. Bruno, "Perceiving occlusion through auditory-visual substitution," *Cogn. Process.*, Vol. 7, pp.128-131, Sep. 2006.

[2] S. Shao, "Mobility aids for the blind," *Electronic Devices for Rehabilitation*., New York: Wiley, pp. 79-100, 1985.

[3] P. Bach-y-Rita, Brain Mechanisms in Sensory Substitution, San Diego, CA: Academic, 1972.

[4] C. Capelle. C. Trullemans, P. Arno, and C. Veraart, "A Real-Time Experimental Prototype for Enhancement of Vision Rehabilitation Using Auditory Substitution," *IEEE Trans. Biomed. Eng.*, Vol. 45, pp. 1279-1293, Oct. 1998.

[5] P. Meijer, "An experimental system for auditory image representations," *IEEE Trans. Biomed. Eng.*, vol. 39, pp. 112-121, Feb. 1992.

[6] P. Arno, A. Vanlierde, E. Streel, M.-C. Wanet-Defalque, S. Sanabria-Bohorquez, and C. Veraart, "Auditory Substitution of Vision: Pattern Recognition by the Blind," *Appl. Cognit. Psychol.*, Vol. 15, pp. 509-519, Sep. 2001.

[7] J.Borenstein, "The NavBelt-A Computerized Muti-Sensor Travel Aid for Active Guidance of the Blind," in *Proc. 5th Annu. Conf. Technology and Persons with Disabilities*, Los Angeles, California, pp. 107-116, Mar. 1990.

[8] Y. Kawai, and F. Tomita, "A Support System for Visually Impaired Persons Using Acoustic Interface-Recognition of 3-D Spatial Information," in *Proc. 16th Intl. Conf. on Pattern Recognition*, Quebec City, Canada, Vol. III, pp. 974-977, Aug.2002.

[9] K. J. Palomäki, H. Tiitinen, V. Makinen, P. J. C. May, and P. Alku, "Spatial processing in human auditory cortex: The effects of 3D, ITD, and ILD stimulation techniques," *Brain Res. Cogn. Brain Res.*, Vol. 24, pp. 364-379, Aug. 2005.

[10] V. R. Algazi, R. O. Duda , and D. M. Thompson, "The CIPIC HRTF database," in *Proc. of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustic*, New Paltz, NY, 2001.