TECNIACUSTICA VALLADOLID·2013

44º CONGRESO ESPAÑOL DE ACÚSTICA
ENCUENTRO IBÉRICO DE ACÚSTICA
EAA EUROPEAN SYMPOSIUM ON ENVIRONMENTAL ACOUSTICS AND NOISE MAPPING

OBJECT DETECTION AND REPRESENTATION PROCEDURE FOR A NAVIGATION DEVICE FOR BLIND PEOPLE

PACS: 43.66.Qp

Larisa Dunai, Ismael Lengua, Guillermo Peris, Beatriz Defez
Centro de Investigación en Tecnologías Gráficas
Universitat Politècnica de València, UPV
Valencia, Spain
E-mail: ladu@upv.es

ABSTRACT

Sound source localization was measured as a function of time by using a navigation device for blind people. The device is based on a 3D-CMOS Time of Flight distance measurement method and an acoustic module. The device detects the obstacles in the front of view and represents them through stereo sounds. The sound was previously convolved with Head Related Transform Functions, measured in the anechoic chamber by using Maximum Length Binary Sequence. The sound level is 60 Db at frequency of 44.1 kHz delivered through stereo headphones. The method of user response was showing with the hand the direction and location of the sound source position. Seven experimental exercises were processed, a single object, two objects a plane wall, etc., where each object was represented by multiple sound sources. The best time results of sounds source localization occurred at localization of one object 0.57 min. The highest time is 12,5 minutes where one column was placed in front of a wall at distance of 1 meter. Sound source localization appears to show the best method for object detection for electronic navigation devices for blind people, due to their hearing abilities.

Keywords—Assisted navigation; blind mobility; range sensor; audio map; computer vision.

RESUMEN

La localización de una fuente de sonido ha sido representada como una función del tiempo mediante el uso de un dispositivo de navegación para las personas ciegas. El dispositivo se basa en un sistema 3D-CMOS utilizando el método de medición de la distancia de vuelo y un módulo acústico. El dispositivo detecta los obstáculos y los representa a través de sonidos estéreo. El sonido fue convolucionado previamente mediante HRTF, medido en cámara anecoica usando “Maximum Length Binary Sequence”. El nivel de sonido es de 60 dB en la

frecuencia de 44,1 kHz suministrado a través de auriculares estéreo. El método de respuesta del usuario se muestra identificando con la mano la dirección y la ubicación de la posición de la fuente de sonido. Para la validación se realizaron siete ejercicios experimentales, el primero un único objeto, dos objetos de un plano de pared, etc, donde cada objeto se representa mediante múltiples fuentes de sonido. Los mejores resultados de origen localización se produjeron en la localización de un objeto 0.57 min. El tiempo máximo es de 12,5 minutos, donde se colocó una columna delante de una pared a una distancia de 1 metro. La localización de la fuente de sonido parece ser el mejor método para la detección de objetos en los dispositivos electrónicos de navegación para invidentes, debido a su gran capacidad de audición.

Keywords—Navegación asistida, movilidad ciegos, mapa de audio, visión ordenador.

I. INTRODUCTION

Information in the environment enables humans and vertebrates to learn about sources that are in many different directions, particularly signals that are outside the detection range of other senses [1]. Sound source localization is inherently important for safety-survival and navigation. Blind people make maximum use of sound not only to know the obstacle presence but also where is and how dangerous is in order to avoid it effectively.

There are over 314 million of blind and partially sighted people in the world from where 39 million are total blind [2]. Blindness is the condition of lacking visual perception due to physiological or neurological factors.

There are several main skills that the blind community requires:

1. live independently and productively
2. communication
3. raise a family
4. have a social life
5. mobility
6. maintain a career- or launch a new one
7. enjoy sports, games

Loss of vision often is accompanied by loss of independence. Visual impaired and, in particular, total blind people are unable to take advantages of different services. They have lack of social interaction, human contact and they are limited in mobility.

Nowadays various techniques are developed for reading and writing for blind community: Braille, talking books, reading machine which convert the printed text into speech or Braille. Also a variety of computer software and hardware such as mobiles, scanners and refreshable Braille display, optical character recognition applications and screen readers, radio reading services etc., help blind community to communicate with the surrounding people, familiars and unfamiliar people.

One of the main necessities of the blind people is the lack of mobility, which become a severe constraint for the person. Blind people find difficulties to travel independently, because they cannot determine their positions and objects location in the surrounding environment. For the sole purpose of getting out a considerable amount of information is required.

Blind Unions help blind users to learn to use various techniques and methods of reading, writing and navigate. Also learn how to improve other body part which will help them to orientate and perceive the surrounding. They learn to make use of the sounds, feelings, temperature, etc. to help them in their habitual life.

Most of blind and partially sighted people learn to use their audition to compensate the lack of vision. Environmental information enables the humans and animals to learn about sources and sounds from the surrounding.

Acoustic information is a primary tool for orientation by blind and partially sighted people, for example, to determine when traffic has actually stopped – rather than when it has been signaled

to stop. And when crossing at an intersection that has no traffic lights, they listen for oncoming traffic to determine when to cross.

During many centuries the “cane” has been the most popular mobility aid system. Despite its importance in the blind community, before 1964, when Russell C. Williams published the “Specifications for the long cane (Typhlocane)”, which helped to establish a long cane model, the used canes lacked of any standards and specifications [3].

Beside the working requirements and system specifications, an ETA system should be designed in order to contain the minimum possible number of accessories (boxes, electronics, helmets and connecting cables, etc.,) in order not to bother and disturb the user. Besides the importance of the system accessories, it is also significant the importance of the techniques for representing the information acquired from the environment.

From a technical point of view, the devices should be designed and developed not only to be light and small, but also reliable and durable and esthetical well designed, to have a high quality and to assure reliability during their operation.

Technically, the ETA systems are based on tree interfaces: the input interface, the processing interface and finally the output interface. The input interfaces acquire the environmental data. They can be classified in: ultrasound, laser, and artificial vision and GPS systems. The processing interface contains the techniques and software for processing all the acquired information and for transforming it into the required data for the output interface. The output interface is, as the previous interfaces, important. The output system represents the model for transmitting the information from the device to the user. It should be as much concise and clear as possible, in order not to confuse and disturb the user.

II. Cognitive Aid System for Blind People (CASBliP)

The main aim of CASBliP is to develop an integrated system able to interpret and manage real world information from different sources of information in order to help blind people to navigate through the environment.

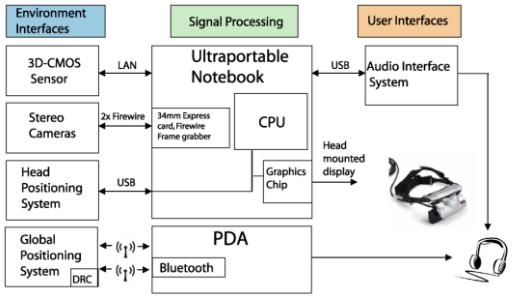

Fig. 1. Overall system schema

A. Depth estimation and image segmentation

In order to estimate the distance of objects from the user in an efficient manner, a stereo grid with two cameras is used since it is generally faster than single camera temporal depth estimation.

The Firewire interface was thus chosen to allow real time frame rates while supporting two cameras. The intrinsic and extrinsic parameters of the two cameras are pre-computed using a classic chart based calibration technique. Sparse depth estimation, e.g. correlation based patch correspondence search and reconstruction, is usually computationally efficient. However, it is not desirable in the application since it often results in isolated regions even though they may belong to a single object which makes it difficult to reproduce by sounds. In recent years, there has been considerable interest in dense depth estimation, e.g. [4]. Several methods were tested, including belief propagation [5], dynamic programming [6], sum of absolute difference with winner-take-all

optimisation [7], and sum of squared differences with iterative aggregation [8].

Although recent comparative studies, such as [4], suggest that scan line based dynamic programming does not perform as well as more global optimization approaches, on our outdoor dataset it appears to be a good trade-offs between computational efficiency and quality (based on subjective comparison, e.g. see Fig. 2).

Fig. 2. First row: a pair of stereo images and depth estimation based on dynamic programming using 1D optimisation; second row: results obtained from belief propagation, sum of absolute difference with winner-take-all optimisation, and sum of squared difference with iterative aggregation.

Note that the present outdoor images are considerably different from those benchmarks widely used in comparative studies. It is very common to contain disparities of up to 60 pixels out of 320 pixels, which is significantly larger than most standard ones. Additionally, the variation of disparities is large, i.e. for most of the frames the disparity covers most levels from 0 to 60. However, due to the nature of 1D optimisation, streaking artefacts inevitably results. A median filter across the scan lines is used to reduce this effect. More advanced approaches, e.g. [9], can be used.

Since the stereo cameras are constantly moving and the scene often contains moving and deforming objects, enforcing temporal consistency does not necessarily improve results. A fusion approach using image segmentation based on the assumption that depth discontinuity often collocates with discontinuity in regional statistics was adopted. Similar ideas have been recently explored, e.g. [10]. However, a post-fusion approach instead of depth estimation from over-segmentation[1] was adopted. Smoothing is performed within each region to avoid smudging across the region boundaries. An example result is shown in Fig. 3. This segmentation is based on graph a cut [11], which offers the potential of multimodal fusion of depth, colour components and sparse optical flow to obtain more coherent segmentation. Mean Shift segmentation [12] in the interest of further efficiency were currently investigated. Additionally, these unsupervised approaches allow the system to be used in different locations without lengthy training.

Fig. 3. Fusion of depth map with image segmentation. 1st row: the original left image and graph cut based segmentation using colour and raw depth information; 2nd row: original depth estimation and result after anisotropic smoothing based on segmentation.

B. Obstacle detection and motion estimation

Objects immediately in front of the user may pose danger and must be detected. This can be

[1] Image segmentation is also required as part of a subsystem in CASBliP, not discussed in this paper, for assistance to partially sighted users. Hence, the computational overhead of fusing depth information and image segmentation is limited

quite efficiently carried out by threshold the dense depth map, with the threshold conveniently obtained based on stereo parameters and distance of interest, e.g. see Fig. 4. Morphological processing can follow to remove isolated regions for the ease of representation by sounds. This information will be used by the sound module to enhance its decision based on the CMOS sensor (described later) which is specifically applied to the detection of immediate objects up to 5m away. Note, all this will be in addition to the information derived by the user with the traditional white cane. The proposed method here is simple and can better warn the user of obstacles above the ground level.

Fig. 4. Obstacle detection based on depth map thresholding.

C. Object detector using depth-maps

The primary aim with respect to generic object detection is to identify objects moving independently in the scene as these are likely to present the greatest danger. This is achieved by tracking a sparse set of feature points, which implicitly label moving objects, and segmenting features which exhibit motion that is not consistent with that generated by the movement of the cameras. Sparse point tracking has previously been applied successfully to segmentation in [13].

The Kanade-Lucas-Tomasi (KLT) [14] tracker is used to generate this sparse set of points in tandem with the depth estimation process. This tracker preferentially annotates high entropy regions. As well as facilitating tracking, this also ensures that values taken from the depth maps in the vicinity of the KLT points are likely to be relatively reliable as it is probable that good correspondences have been achieved for these regions.

Points corresponding to independently moving objects are segmented using MLESAC [15], based on the assumption that apparent movement generated by camera egomotion is dominant. Specifically, homography is repeatedly used to deduce a provisional model based on the trajectory of four randomly selected points over a sliding temporal window and the most likely model retained. Outliers generally correspond to independently moving objects (but sometimes regions close to the cameras are apparently moving due to the 'parallax effect').

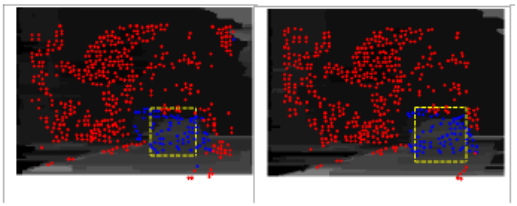

Fig. 5. Object segmentation using depth maps and sparse motion information.

Bounding boxes are fitted iteratively to the segmented points under the assumption that independently moving objects are of fixed size and at different depths in the scene. Firstly, segmented points are aligned with depth maps to ascertain depths for moving object annotation. The mode depth of these points is used to scale a bounding box which is robustly fit to the segmented points, such that number of inliers is maximised. To segment more than one object, the bounding box algorithm is reapplied to the 'bounding box outliers' produced in the previous iteration. This needs to be done judiciously, as these outliers may be misclassified background points that are distributed disparately in the image. However, if a bimodal (or indeed multi-modal) distribution of depths is present, objects will be sequentially segmented in terms of how numerously they are annotated at a consistent depth. Fig. 5 illustrates an example of the proposed method. The left image shows the KLT points segmented into dominant motion (red)

and outliers (blue). The right image shows the depth map aligned with these segmented points and a bounding box fitted as described above.

In addition to fitting a bounding box to the annotation, aligning KLT points with depth maps can generate relative velocity estimates as motion in the image plane may be scaled according to the moving objects distance to that plane. Furthermore, it could potentially allow the system to determine if objects are approaching the user. Figure 6 shows every fourth frame of the segmented foreground region and bounding box fitted.

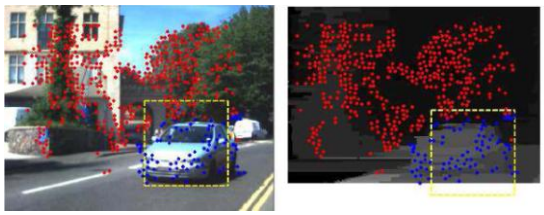

Fig. 6. Every 4th frame of car approaching the viewer

D. Integration with the inertial sensor

As mentioned earlier, the camera is constantly moving which makes it very difficult to estimate object motion, particularly in real time. In order to more efficiently and accurately carry out camera motion estimation, an integrated inertial system, which includes a 3D gyro sensor, a 3D accelerometer, and a magnetometer, is used with the stereo system. This involves both hardware level and software level integration. The inertial sensor is fixed to the stereo rig and carefully placed so that the centre of this sensor is as close to that of the left camera as possible. It primarily acts to provide rotation angles and accelerations for the stereo cameras.

An inertial sensor motion with a 15 element state vector zI =(Θ, Ω, x, v, a)T where Θ is the orientation of the inertial sensor with respect to the world (defined by magnetic and gravitational fields), Ω is the angular velocity, and x, v, a are the position, velocity, and acceleration of the sensor with respect to the world is presented. The readings from the inertial sensor are synchronised with the image acquisition. We also developed an automated method to calibrate the inertial sensor and stereo cameras with the aid of a standard calibration chart, i.e. computing the rotation matrix between the inertial coordinates and the left camera coordinates. The translation between these two can be ignored due to the fact that they are very close to each other [16]. The pre-calibrated stereo cameras are aimed at a calibration chart from various angles and distances. Stereo image pairs are automatically selected from this calibration sequence subject to the requirements that the charts are successfully detected and the cameras are relatively stationary to avoid motion blur, which can be read from the inertial sensor rotational and translational measurements. Thus, for each pose transition we have estimates from the stereo camera, ωC, and measurements from the inertial sensor, ωI. The rotation motion relationship between the two coordinates can be derived as ωC = RICωI , which is solved using least squares optimisation. We are currently integrating this into motion estimation and developing a Kalman filter to reduce the signal noise, which can arise from the motion of the user and linearization error.

E. 3D CMOS sensor

The second component of environment scanning system is based on the 3D-CMOS line sensor with 64 pixels. The system was previously designed by SIEMENS and patented for pedestrians.

The measurement principle is based on Time of Flight (ToF) measurement of pulse-modulated laser light utilising a high-speed photosensitive CMOS sensor and infrared laser pulse illumination. The analogue signals of several laser pulses are averaged on chip to reduce the required laser power and also to increase measurement accuracy. As described by Mengel et al. [17], the measure distance range d for the individual pixel is recovered by:

$$d = \frac{c}{2} \times \left(T_1 - T_w \frac{U_1}{U_2} \right) \quad (1)$$

where T1 is the short integration time, Tw is the pulse width of the laser, U1 is the measured sensor signal at integration time T1, and U2 is the sensor signal measured for integration time T2.

Hence, the camera image contains concurrent distance and intensity information for every single image pixel in real-time in a measurement range of up to 5m, aiming at the detection and localisation of objects (e.g. obstacles). Due to the applied sensor principle, the images will be virtually independent of the prevailing background illumination (i.e. darkness or bright sunshine). A fully solid state micro system with embedded Floating Point Gate Array (FPGA) processing and special optical and mechanical design has been developed, manufactured and integrated into a pair of spectacles as shown in Figure 7.

The essential specification of the 3D-CMOS sensor is summarized in Table 1.

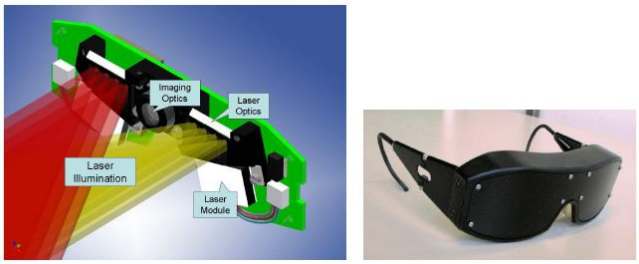

Fig. 7. Sensor design and placement for aesthetic wear

Attribute	Specification	Attribute	Specification
Chip technology	Standard CMOS technology	Field of view	60º x 2º
Number of pixels	64 pixel line	Measurement range	0.5m -5m
Target reflectivity	5% -100%	Measurement time	< 50 ms
Measurement accuracy	< 3% of object distance		

Table I. *3D-CMOS specifications*

F. GPS system

To test the system an area in a University and geo-referenced it with an high-precision D-GPS system, obtaining a map with ±1cm accuracy was used. This data set was then used to compare the data generated by the proposed GPS system. The yellow path shown in Figure 8 represents the walk path that the user was required to follow; the reference path was 338m long and 0.7m wide. The red path signifies the data produced by the GPS. For this open urban path we calculated an uncertainty on the position averaging <1.5m over 3000 data points. The system will next be trialled with visually impaired volunteers in city spaces in collaboration with the Italian Institue of the Blind ”F.Cavazza”, another project partner.

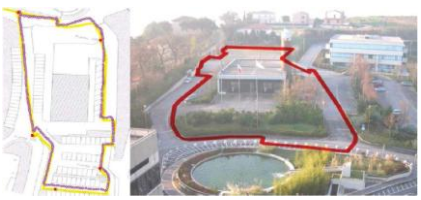

Fig. 8. Tests on full open area (GPS + EGNOS + inertial sensor)

G. Audio interface

A series of experimental tests were carried out to get a better understanding of the role of several significant acoustic parameters in the generation of a spatial auditory image of the 3D scene: the interclick interval, the sound reverberation level and the tonal colour of the click sound. By way of illustration we present the results of a preliminary study on the effects of using or not a tonal codification in the azimuth dimension for both perceiving and localizing separated sound sources inside a field of view of 60. Two experienced subjects indicate the number of perceived virtual sound sources (2, 3 or 4) which are presented in 3 conditions: with no azimuth tonal coding (monocolour click condition), both pre-training and after a short period of training, and with an azimuth tonal coding (bitonal condition). The results in Fig. 11 show that the subjects have almost no problems to identify the presence of two virtual sound sources (each one located at the extremes of the field of view) in any condition, but a clear improvement occurs when trying to identify the presence of three or four sounding objects in the bitonal condition (100% detection), versus the monotonal one. These results suggest that using different sounds for codifying different objects which are present in the scene improve the quality of the corresponding spatial auditory image created in the subject. This result is very promising for designing the strategy for the auditory representation of a larger 3D scene.

A portable audio module has been developed which is able to receive the sensory data and translate them into the appropriate acoustic representation. It keeps in memory the set of sounds to be delivered, and it is able to generate a 3D sound map by adding and reproducing in real time the selected sounds. We use a basic learning protocol for initiating the user in the use of the device. The main objective here is in helping the user to acquire the externalization effect, i.e., the perceptual illusion that the sounds are coming from an external point to the user, but not from the headphones. A series of studies to measure the user's performance in different Orientation and Mobility (O+M) tasks are being currently carried out on a wide sample of blind participants. The first preliminary results show that a majority of the users achieve a correct performance in these initial O+M tasks, which suggest that they are adequately experiencing the expected spatial auditory images that correspond to the presented scenarios.

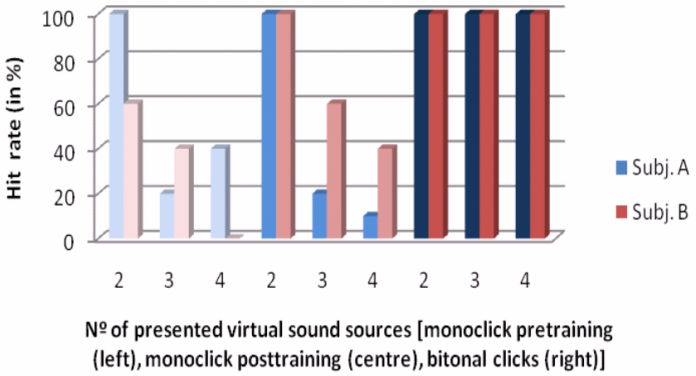

Fig. 11. Role of the azimuth tonal coding on the perception of separate sound sources

After long testing periods in the laboratory, the acoustic sounds were implemented into the navigation device CASBliP. Previously, the device was tested in laboratory conditions with twenty blind people. Seven simple exercises: detection of one obstacle, detection of two obstacles, detection of the space between two obstacles, detection of an obstacle in front of another obstacle were carried out. Each participant in the tests should detect through acoustic sounds delivered by the device through headphones and show with the hands the object position and its volume.

In Fig.12 are presented the results obtained during the seven exercises. The participants were divided in two groups: ten in the group A and ten in the group B. The group A repeated three times the same experiment (blue color). For the exercise 1 the group A obtained great improvement from 1,19 min to 0,57 min and the standard deviation of 0,31. The group B have the average time 2,15 min. On the second exercise the group A obtained 2,01min average time with a standard deviation 0,59 and the group B 4,40 min. The maximum time for the first exercise on the group A is 2,17 min when in the group B is 3,30min. For the exercise 2 the maximum time for

the group A is 4,57 and the group B is 5 min. In the group A was perceived great improvements after each trial. For example if the first trial the registered time is 6,24 for the exercise 6 at the second trial the time is 4,04min and at the third trial is 2,44min.

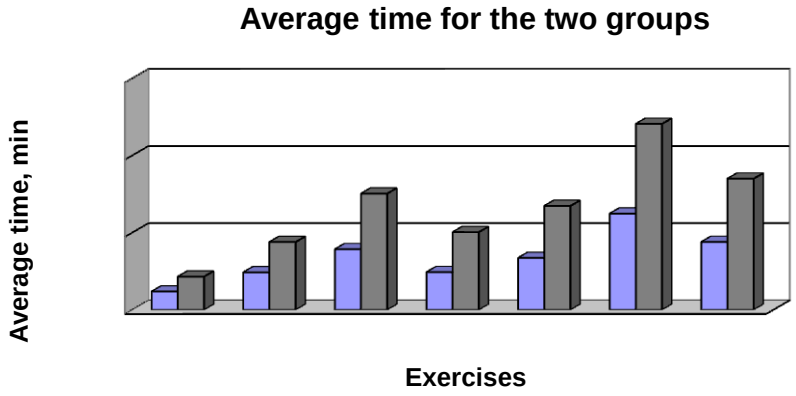

Fig. 12. Time average of object localization for seven exercises

Regarding results obtained on the navigation by using acoustic sounds, we can mention that the walking time improved almost 40% from one trial to another. It means that the blind people make great use of acoustical sounds and their hearing, improving their perception ability. Also, it demonstrate that acoustic sounds are useful for human guided devices.

III. CONCLUSION

It has been also proven that with the developed system, the blind people is able to travel confidently and safety. Due to the fact that the system gives information on the environment comprised between 0 and 15m within a range of 32º relative to the right and left side of the user, the system gives more information than the white cane or the existing ETA systems, which are constricted to distances up to 6m.

A system based on a CMOS Time of Flight laser is used for detection of the objects between 1 and 5m, giving a measurement error lower than 1% distance for 100% target. Also the system can be used with decreasing resolution and accuracy in distances over 5 meters.

An object detection system based on Stereo Cameras can be used in parallel or individually. It shows that it is possible to represent the real environment using the depth map method extracted from the stereo vision, to extract the objects, to classify them and to extract the free paths.

It has been also proven that using spatial acoustic sounds, the processing and delivering time is not long, since the used spatial sound has 2048 samples. Due to the short acoustic sounds used on the system, the representation of the objects detected by the system is delivered in real time. It is not necessary to inform the user about the objects located in front of view one by one; the system represents the objects at different peach, depending on their distance and type, thus the user perceives the whole image of the environment being able to take decisions.

ACKNOWLEDGMENT

The work has been supported by the project Nº 2062 “Desarrollo de un sistema de entrenamiento acústico virtual para localización de sonidos espaciales para personas invidentes”, granted by Universitat Politècnica de València

REFERENCES

[1] R.R. Fay, A.N. Popper. “Introduction to sound source localization,” Springer Handbook of Auditory research, sound source localization, pp. 1-5, 2005

[2] WBU “White cane safety day,” World Blind Union, Press release, October, Canada , 2009.

[3] W. Farmer “Mobility devicesa,” Bulletin of Prosthetic Research, pp-47-118, 1978.

[4] D. Scharstein and R. Szeliski. “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” Int. J. Comput. Vision, 47(1-3), pp.7–42, 2002.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. “Efficient belief propagation for early vision,” Int. J. Comput. Vision, 70(1), pp.41–54, 2006.

[6] S. Birchfield and C. Tomasi. “Depth discontinuities by pixel-to-pixel stereo,” Int. J. Comput. Vision, 35(3), pp.269–293, 1999.

[7] T. Kanade. “Development of a video-rate stereo machine,” In Image Understanding Workshop, pp. 549–9557, 1994.

[8] C. Zitnick and T. Kanade. „A cooperative algorithm for stereo matching and occlusion detection,” IEEE Trans. PAMI, 22(7), pp.675–684, 2000.

[9] A. Bobick and S. Intille. “Large occlusion stereo,” Int. J. Comput. Vision, 33(3), pp.181–200, 1999.

[10] C. Zitnick and S. Kang. “Stereo for image-based rendering using image over-segmentation,” Int. J. Comput. Vision, 75, pp.49–65, October 2007.

[11] P. F. Felzenszwalb and D. P. Huttenlocher. “Efficient graph-based image segmentation,” Int. J. Comput. Vision, 59(2), pp.167–181, 2004.

[12] D.‘Comaniciu and P. Meer. “Mean shift: a robust approach toward feature space analysis,” IEEE Trans. PAMI, 24(5), pp.603–619, 2002.

[13] S. Hannuna. “Quadruped Gait Detection in Low Quality Wildlife Video,” PhD thesis, University of Bristol, 2007.

[14] J. Shi and C. Tomasi. “Good features to track,” In CVPR 94, June 1994.

[15] P. Torr and A. Zisserman. “Mlesac: A new robust estimator with application to estimating image geometry,” Comp. Vis. and Image Understanding, 78, pp.138–156, 2000.

[16] J. Lobo and J. Dias. “Relative pose calibration between visual and inertial sensors,” Internation Journal of Robotics Research, 26(6), pp.561– 575, 2007.

[17] P. Mengel, G. Doemens, and L. Listl. “Fast range imaging by CMOS sensor array through multiple double short time integration (mdsi),” ICIP, pp. 169–172, 2001.