# A Cross-Modal Electronic Travel Aid Device

F. Fontana, A. Fusiello, M. Gobbi, V. Murino,
D. Rocchesso, L. Sartor, and A. Panuccio

Dipartimento di Informatica, University of Verona,
Ca' Vignal 2, Strada Le Grazie 15, 37134 Verona, Italy,
{fontana,fusiello,murino,rocchesso,panuccio}@sci.univr.it

**Abstract.** This paper describes the design of an Electronic Travel Aid device, that will enable blind individuals to "see the world with their ears." A wearable prototype will be assembled using low-cost hardware: earphones, sunglasses fitted with two CMOS micro cameras, and a palmtop computer. Currently, the system is able to detect the light spot produced by a laser pointer, compute its angular position and depth, and generate a correspondent sound providing the auditory cues for the perception of the position and distance of the pointed surface. In this way the blind person can use a common pointer as a replacement of the cane.

## 1 Introduction

Electronic Travel Aids (ETA) for visually impaired individuals aim at conveying information in a way that the user can reconstruct a scenario alternative to the visual one, but having similar informative characteristics, such that the impaired person can make use of at least part of the information that sighted people normally use to experience the world. This information is typically conveyed via the haptic and auditory channels.

Since the mid-seventies, a number of vision substitution devices have been proposed which claim to convert visual into auditory or haptic information with the aim of being mobility aids for blind people [9,6]. Early examples of this approach are the Laser-Cane and the Ultra Sonic Torch [12]. These devices are all based on beams of ultrasonic sound which produces either an audio or tactile stimulus in response to nearby pointed objects, proportional to the proximity of the detected object. In the VIDET project a device converting the information of object distance into haptic cues has been developed [13].

As compared to visual information, sound does not rely on focused attention of the human subject, it is pervasive in a 3-D environment. However, it is difficult to design sounds in such a way that they become effective information carriers. When one aims at using sounds to display spatial features of objects, it should be considered that, while vision is best suited for perceiving attributes of light modifiers (surfaces, fluids), audition is best suited for perceiving attributes of sound sources, regardless of where they are located [10].

F. Paternò (Ed.): Mobile HCI 2002, LNCS 2411, pp. 393–397, 2002.

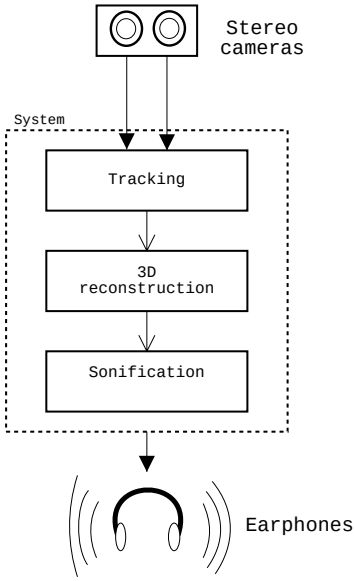


**Fig. 1.** Overall system architecture. The laser spot is tracked, then its 3D position is computed and sonified.

In this research, we export to ETAs the idea of augmenting the humans' visual perception using acoustic stimuli.
This should lead to novel interface designs and, in particular, to new strategies for the designs of ETAs and other aids for blind people. On the visual side, we have to deal with methods that allow recognition and positioning of the objects that are present in the scene, and, on the auditory side, with techniques that allow blind people to experience the presence of such objects through synthetic sounds.
In spite of increasing power of wearable computers, there is relatively little research on the integration of audio and visual cues for environment perception, although the potential benefits can be very large in the field of the human-computer interaction (HCI). This paper describes a system prototype where such issues are applied.
The system is composed by earphones, two CMOS micro cameras (a stereo head), and a computer (see Fig. 1). The system is able to detect the light spot produced by a laser pointer, compute its depth, and provide suitable spatialized sounds to a blind person.

## 2 Visual Analysis

The human visual system gains information about the depth basing on differences (disparity) between the images formed on the left and right retina. This process, called stereopsys [7,4], can be replicated by a computer, at a rate of some frames per second, using a camera pair. A well-known problem in computational stereopsys is finding *correspondences*, i.e., finding which points in the left and right images are projections of the same scene point (a *conjugate pair*). Using a laser pointer makes this task trivial, provided that we are able to locate the laser spot in the two images.

In the following the processing applied to each image (left and right) in order to locate the target (i.e. the centroid of the blob produced by the laser pointer) in each image will be briefly described.

The visual analysis starts by applying a red band-pass optical filter on both cameras, in order to enhance the image of the red laser spot. Then images are smoothed in time, by averaging pixels at the same position in the preceding $m$ frames (we used $m = 2$). Only pixels whose difference is less than a given threshold are averaged, thereby obtaining a remarkable increment of the stability of the image, without introducing motion blur. Images are binarized with an

automatic threshold and a size filter is applied, which discards the connected components (or blobs) whose size is outside a given range. In order to make the spot detection more robust, we impose the epipolar constraint [8] on the surviving blobs; only those satisfying the constraint both left to right and right to left are kept as candidate targets.

The number of candidate targets is used as a feedback to set the binarization threshold, which is varied using a dichotomic search until the candidate targets are more than zero and less than few units (typically 5). If a minimum threshold value is reached no targets are selected, presumably because the laser spot is not visible. To increase the precision of tracking, a forecast of the position of the pointer in the image is used, assuming a constant speed model [1]. The predicted position will be used to match the current target to one of the target candidates found by the preceding elaborations. In the one-target case the most common choice is to take the closest candidate, as stated by the *Nearest Neighbor Data Association* algorithm [1].

Finally, triangulation recovers the full 3-D coordinates of the laser spot, using the disparity of the targets in the two stereo images, the internal parameters of the two cameras, and the geometry of the stereo setting (obtained from calibration [8]). Each conjugate pair ideally defines two rays that intersect in a point of the space. In real world, for errors in measurement of point positions and in calibration of the stereo camera, the rays don't always intersect, so the triangulation computes the coordinates of a 3-D point nearest to both rays [15].

## 3 Sonification

Sounds that give directional cues are said to be "spatialized" [2]. In other words, they provide distance, azimuth and elevation cues to the listener with respect to the position of an object, assumed to be a sound source. Despite all research done (see [3]), we have only partial knowledge about how we determine the position of a sound source, and even less about how to recreate this effect. In everyday listening, hearing a monophonic source gives not only information about the source position, but also characteristics about the environment, such as source dimension and shape of the room.

The characteristics that a sound acquires during its path from source to the listener's ear canal determine the spatial cues that are conveyed to the listener by the sound. Our model adopts a versatile *structural* model for the rendering of the azimuth [5]. A model providing distance cues through loudness control and reverberation effects is in an advanced stage of development. Further studies are needed to convey reliable elevation cues to the listener.

Once the model for the azimuth [5] has been adapted and fine-tuned to our application, we need a new, independent model capable of rendering distances. In this way we can reasonably think to implement these two models independently in the system, typically in the form of two filter networks in series, the former accounting for distance, the latter for azimuth.

The listening environment we will consider is the interior of a square cavity having the aspect of a long tube, sized 8×0.4×0.4 meters. The internal surfaces of the tube are modeled to exhibit natural absorption properties against the incident sound pressure waves. The surfaces located at the two edges are modeled to behave as *total* absorbers, to avoid the creation of echos inside the tube [11]. It seems reasonable to think that, although quite artificial, this listening context conveys sounds that acquire noticeable spatial cues during their path from the source to the listener.

Given the peculiar geometry of the resonator, these cues mainly account for distance. The square tube has been modeled by means of finite-difference schemes [14]. In particular, a *wave*-based formulation of the finite-difference scheme has been used, known as the Waveguide Mesh, that makes use of the wave decomposition of a pressure signal into its wave components [16]. Waveguide Digital Filters provide wall absorption over the internal surfaces of the tube [17].

## 4 Preliminary Results and Conclusions

Tests show that the effectiveness of laser tracking depends mainly on how "visible" the laser is inside the captured image. Therefore, if the scene is (locally) brighter than the laser spot or if the laser points too far away from the camera, then problems arise about the stability of laser tracking, because the signal to noise ratio becomes too low. Better results are obtained indoor, with a 4 mW laser source, and within a depth range of 1 to 6 meters. The power of the laser, the camera gains and the resolution are critical parameters.

We evaluated the subjective effectiveness of the sonification informally, by asking some volunteers to use the system and report their impressions. The overall result have been satisfactory, with some problems related to the lack of perception of elevation. Further usability tests of this device are planned, with a group of both sighted and visually-impaired subjects. Precision and latency of the system will be measured.

The proposed system represents a first attempt to integrate vision and sound in an HCI application and, for this reason, it is open to further improvements. The next version of the system will determine the three-dimensional structure of a scene by dense stereopsys, then segments the objects (or surfaces patches) contained therein basing on depth. From that information, it will synthesizes spatialized sounds that convey information about the scene.

We are also migrating to a new architecture based on a iPAQ 3760 PocketPC under Linux with a digital color stereo head by Videre Design. This new system will provide a better resolution which can be exploited for the detection of colors and textures. Moreover, future developments of this prototype will be devoted to the design of alternative sound models in order to provide better distance cues, together with elevation. Another line of improvement lies in the coupling of sonic and visual information coming from objects surfaces, e.g., to acoustically render material or texture, so to enrich the listener's experience about the surrounding environment.

**Acknowledgments**

This work is part of the "Sounding Landscape" (SoL) project, supported by Hewlett-Packard under the Philantropic Programme.

## References

1. Y. Bar-Shalom and T. Fortmann. *Tracking and Data Association*. AP, 1988.
2. Durand R. Begault. *3D Sound for virtual reality and multimedia*. AP Professional, 955 Massachusetts Avenue, Cambridge, 1994.
3. J. Blauert. *Spatial Hearing: the Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA, 1983.
4. R. C. Bolles, H. H. Baker, and M. J. Hannah. The JISCT stereo evaluation. In *Proceedings of the Image Understanding Workshop*, pages 263–274, Washington, DC, April 1993. ARPA, Morgan Kaufmann.
5. C. P. Brown and R. O. Duda. A structural model for binaural sound synthesis. 6(5):476–488, September 1998.
6. Jr S. A. Dallas. Sound pattern generator. WIPO Patent No. WO82/00395., 1980.
7. U. R. Dhond and J. K. Aggarwal. Structure from stereo – a review. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6):1489–1510, Nov/Dec 1989.
8. O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. The MIT Press, Cambridge, MA, 1993.
9. R. Fish. An audio display for the blind. *IEEE Trans. Biomed. Eng.*, 23(2), 1976.
10. M. Kubovy and D.Van Valkenburg. Auditory and visual objects. *Cognition*, 80:97–126, 2001.
11. Heinrich Kuttruff. *Room Acoustics*. Elsevier Science, Essex, England, 1991.
12. Kay L. Air sonar with acoustical display of spatial information. In *Animal Sonar System*, pages 769–816, New York, 1980.
13. The LAR-DEIS Videt Project. University of Bologna - Italy. Available at URL http://www.lar.deis.unibo.it/activities/videt.
14. J. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Wadsworth & Brooks, Pacific Grove, CA, 1989.
15. E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice-Hall, 1998.
16. Scott A. Van Duyne and Julius O. Smith. Physical modeling with the 2-D digital waveguide mesh. pages 40–47, Tokyo, Japan, 1993. ICMA.
17. Scott A. Van Duyne and Julius O. Smith. A Simplified Approach to Modeling Dispersion Caused by Stiffness in Strings and Plates. pages 407–410, Aarhus, Denmark, September 1994. ICMA.