# A Multimodal Smartphone Interface for Active Perception by Visually Impaired

Nicola Bellotto
School of Computer Science, University of Lincoln
Lincoln, United Kingdom
nbellotto@lincoln.ac.uk

***Abstract*—The diffuse availability of mobile devices, such as smartphones and tablets, has the potential to bring substantial benefits to the people with sensory impairments. The solution proposed in this paper is part of an ongoing effort to create an accurate obstacle and hazard detector for the visually impaired, which is embedded in a hand-held device. In particular, it presents a proof of concept for a multimodal interface to control the orientation of a smartphone's camera, while being held by a person, using a combination of vocal messages, 3D sounds and vibrations. The solution, which is to be evaluated experimentally by users, will enable further research in the area of active vision with human-in-the-loop, with potential application to mobile assistive devices for indoor navigation of visually impaired people.**

***Index Terms*—Mobile assistive technology, human-in-the-loop, active perception, pervasive computing, 3D audio.**

## I. Introduction

This work is motivated by research interests in the applicability of concepts from the well-known active perception paradigm [1], in which cameras controlled by electro-mechanical actuators have been so far the predominant subject of study [2], to the case of systems with *human-in-the-loop*, where a person holds the sensor and moves it according to automatically generated instructions. The proposed research finds application in the field of mobile assistive technologies to aid people with sensory impairments. The system here considered, indeed, is a smartphone with camera, inertial sensors and audio/tactile interfaces helping a visually impaired (VI) person to navigate in indoor environments, detecting important landmarks for localization and possible obstacles or hazards. In this work, in particular, a multimodal interface is developed, which combines sound, vibrations and vocal message as a means to "control" the orientation of a smartphone camera being held by a person.

As highlighted in [3], modern mobile devices are excellent tools for assisting people with VI. Recent advances in computer vision, in particular, can be exploited to create new assistive devices improving their quality of life. Several solutions have been proposed in the past, for indoor localization and obstacle detection, based on wearable systems that combines GPS, wireless and ultrasound devices [4], or stereo vision [5]. In some cases, cameras have been mounted on wearable pan-tilt units [6] and, although originally developed for applications of augmented reality, one can see their potential for guiding a VI person. More recently, Electronic Travel Aids (ETAs) have been developed based on the ever growing number of tablets and smartphones, for example to implement landmark-based localization using vision [7] or global positioning by means of GPS, compass and inertial sensors, which most mobile devices are now provided with [8]. The VI using these localization systems, however, often require also a separate white cane for obstacle detection. Moreover, in case of non-vision-based systems, the users are still left with the "last 10 meters problem", that is, the impossibility to localize a particular object or passage nearby (e.g. exit door).

Most of the previous research in computer vision for mobile assistive devices has concentrated on the challenging tasks of feature detection, localization and object identification, often communicating the respective information to the user via simple vocal messages, sounds or vibrations [7], [9], [10]. Rather little importance have been given though to the communication aspect, and to the way this information is presented to the VI user. A good interface, however, is essential for such assistive devices, in particular if the latter are active perception systems designed to guide human behaviours.

An attempt to convert images to sound have been investigated by using *The vOICe* application [11]. The whole scene is continuously scanned and converted into "sound images" for the VI. Given the richness of the information, the system requires long periods of training in order to deal with the inherent cognitive load. Some other works, instead, have used 3D sounds to localize particular visual features [12], [13]. Most of these systems, however, rely on the assumption that the camera is oriented like the user's head, which facilitates the spatial representation of sound sources. Unfortunately, in case of a hand-held smartphone, such assumption does not hold.

To enable further research in this area, a new multimodal interface for smartphones, combining vocal messages, 3D sounds and vibrations, has been developed to guide the pointing actions of VI people with an acceptable degree of accuracy and responsiveness, without overloading the cognitive capabilities of the user during basic navigation tasks inside buildings.

The reminder of the paper is as follows: Sec. II introduces the general context of this research with an overview of the whole perception system, from which the problem of human control is then extracted and analysed in Sec. III. The implementation details on a smartphone are explained in Sec. IV. Finally, Sec. V concludes the paper analysing pro ad cons of the proposed approach, and discussing steps towards future research in active perception for mobile assistive devices.

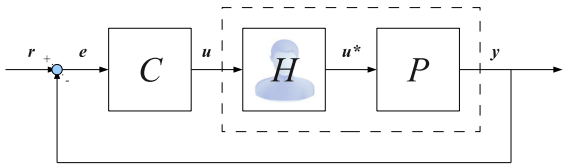


Fig. 1. Feedback configuration with human-in-the-loop.

## II. Active Perception with Human-in-the-Loop

The system here considered is a navigation aid tool that uses vision to detect obstacles in the environments (or landmarks for localization). It can be formulated as a classic problem of active perception, where the objective is to find models and control strategies to allow or facilitate the execution, in this case, of a mobility task [1]. In particular, the processes involved can be represented by a closed-loop system, in which the feedback from the smartphone sensors and the vision algorithms are converted into "control signals" for the user to execute. The goal is to orient the smartphone's camera towards particular objects or features in the environment, the position of which serves as reference for the system.

Fig. 1 illustrates the system with a simplified diagram. The input $r$ is the reference provided, for example, by an obstacle detection or a localization algorithm, giving the direction of a visual target the camera should be pointing at. The error $e$ between the reference and the actual orientation $y$ of the camera is used by the controller $C$ to generate the control signal $u$. It is at this point that the classic active vision paradigm [2] differs from current system: while the former is generally concerned with the optimal control of some electro-mechanical device that regulates the internal and/or external camera's parameters (e.g. position, orientation, focal length, etc.), the latter tries to control the output of the whole human-camera subsystem, illustrated in the figure by block $H$ and $P$ respectively. Using a simple analogy, the user holding the camera corresponds to the typical pan-tilt unit or the mobile robot often encountered in previous active vision systems [14]. The internal signal $u^*$ can be thought as the torque applied by the human to the smartphone, which changes its direction and the orientation of the scene observed by the camera.

One of the problems in dealing with such a system, featuring human-in-the-loop, is to define a suitable form of control signal $u$ that, in combination with the policy defined by $C$, optimizes the system response according to some given criteria, for example in terms of reaction time or accuracy in following the variations of the reference $r$. The work here presented is not intended to provide a mathematical solution to the problem, or an in-depth analysis of the cognitive processes involved in this kind of human-machine interfaces. Instead, it explores a possible way to convey the information that needs to be transmitted from the control algorithm $C$ to the user, i.e. to define a possible "signal" $u$ that, at least in most of the cases, can be interpreted correctly by the person within a reasonable time.

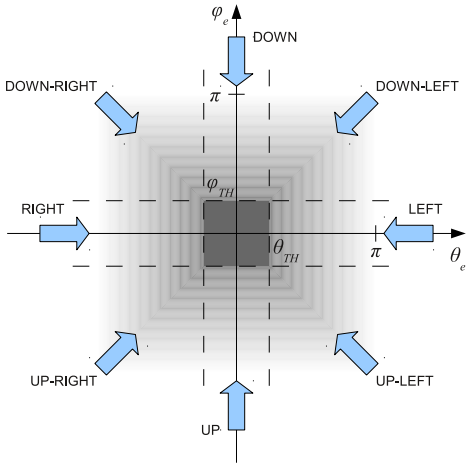


Fig. 2. Mapping of the reference error $e$ to the control signal $u$.

## III. Multimodal Control Interface

Given the reference direction $r = [\theta_r, \phi_r]^T$, where $\theta_r$ and $\phi_r$ are, respectively, the azimuth and zenith of the hypothetical visual feature to be observed, and the current orientation $y = [\theta_y, \phi_y]^T$ of the camera, the input for the controller $C$ is the error $e = r - y = [\theta_e, \phi_e]^T$. Three different levels of feedback are then generated by $C$ as control signal $u$ for the user $H$: a) vocal messages; b) 3D sounds; and c) vibrations.

The mapping from $e$ to $u$ is shown in Fig. 2. The axes of the graph represent the respective angular quantities $\theta_e$ and $\phi_e$, the origin being the desired orientation of the mobile camera (i.e. $\theta_e = \phi_e = 0$). It is divided in eight slices, each one corresponding to a vocal message that indicates the direction towards the origin ("LEFT", "DOWN-LEFT", "DOWN", etc.) when the current $e$ falls within it. Depending on the distance of $e$ from the origin, a sound of varying pitch and gain is also generated. This is represented by the grey shaded area, densely coloured towards the centre and fading out as the distance from the origin increases. Finally, the small darker square in the middle specifies the region where both $|\theta_e|$ and $|\phi_e|$ are below a given threshold (that is, when the camera's orientation is closer enough to the desired one), in which case a short vibration of the smartphone is generated. Each one of these three modalities is discussed in detail below.

### A. Vocal Messages

Two thresholds, $\theta_{TH}$ and $\phi_{TH}$, are defined to check whether the current orientation of the camera is close enough to the desired value. Every time a new measurement is available, if the azimuth and/or the zenith of the camera differs from the reference of more than the threshold (i.e. $|\theta_e| > \theta_{TH}$ and/or $|\phi_e| > \phi_{TH}$), then a vocal message is generated by the device, according to the mapping shown in Fig. 2. So, for example, if only the azimuth error $\theta_e > \theta_{TH}$, while the zenith $|\phi_e| \leq \phi_{TH}$, then the system generates a vocal instruction saying "Point LEFT". If also the zenith $\phi_e > \phi_{TH}$, instead, the system says "Point DOWN-LEFT". If both the angle errors are within the thresholds, no message is produced.

Note that the direction indicated by these messages is relative to the current camera's orientation, not to the person's body or to an absolute world frame of reference. Also, in order not to overload the user with information, a vocal message is generated only the first time a new reference is provided to the system (and if any angle error is above the threshold, of course). After that, the user will follow the direction of a 3D sound to adjust the orientation of the device. The latter is described in detail next.

### B. 3D Sounds

Current audio technology offers the possibility to *spazialize* different sound sources in a 3D virtual environment [15]. Although with some limitations, normal stereo-headphones are sufficient to give the user the impression that these sounds are located at particular distances and directions in space, as previous studies have shown [16]. Similar technologies have been previously applied to improve the accessibility to computer devices or aid the navigation of VI people [12], [13]. However, these solutions are usually based on devices fixed to the head or the body of the user, which makes an important difference in terms of perceptual flexibility and representation, as discussed in Section IV-B.

In the current system, 3D audio is used as a means to indicate the direction of reference for the camera, which has to be oriented by the user. In practice, the angles $\theta_e$ are $\phi_e$ are used to project a particular sound source into the smartphone's 3D audio environment, and from there to its headphones. The user is then guided to the proper direction by pointing the device towards the sound source, so that the latter appears to be frontally located.

To help the user in moving towards the desired direction, an exponential function has been applied to both the volume and the pitch of the sound: when $|\theta_e|$ and $|\phi_e|$ are below the usual thresholds, the sound is loud and high-pitched; as soon as the angular errors increase, it gets quieter and lower in tone. This exponential change is illustrated in Fig. 2 by the variation of grey shade, denser around the origin (i.e. maximum volume and pitch) and almost disappearing towards $\pm\pi$ (i.e. minimum volume and pitch).

### C. Vibrations

The last interaction modality to communicate with the user is a simple smartphone vibration, which signals the desired orientation has been reached (i.e. $|\theta_e| < \theta_{TH}$ and $|\phi_e| < \phi_{TH}$) and the respective image or visual feature acquired by the system. As for the voice case, only a single vibration is generated. In practice, while a vocal message starts the target acquisition process, the vibration indicates its successful end.

## IV. System Implementation

The system has been implemented on a Nexus 4 (Fig. 3), which is a quad-core smartphone with 2GB of RAM and standard features found in many other models, including an 8MP camera, accelerometers, gyroscopes, and vibration alert. The operating system is a recent version of Android,


Fig. 3. Smartphone implementing the multimodal interface for active perception.

for which several libraries exist to develop dedicated mobile applications (i.e. Android SDK and NDK for native-code C/C++ development [17]). The smartphone is provided with standard headphones, which are necessary for implementing the 3D sound interface. The Text-to-Speech (TTS) used is based on the IVONA engine [18], which provides slightly more human-like voices compared to the standard Android TTS.

The reference angles $[\theta_r, \phi_r]$ of the hypothetical visual targets are currently fixed, as they will be provided by a complementary localization and/or object identification algorithm in future extensions of the system. The orientation of the camera, $[\theta_y, \phi_y]$, is given by the inertial sensors of the smartphone, and is therefore affected by cumulative errors. The latter, however, are negligible for the purpose of initial testing. Once included in the final system, they will be eventually corrected by the visual-inertial localization algorithm.

Given the field of view of the frontal camera, which is approximately 40°, the thresholds around visual targets have been set to $\theta_{TH} = \phi_{TH} = 15°$. Even in the worst case, it is therefore guaranteed that, once the smartphone vibrates, the visual target of reference lies within the current camera view (of course, provided the target is a point-like image features, or some object not too big and not too close to the camera). The thresholds can be adjusted depending on the particular application and smartphone model.

### A. OpenAL

Perhaps the most interesting component of the system interface is the 3D sound spatialization discussed in Section III-B. To implement this, an Android porting of the popular OpenAL library has been adopted, which provides a set of methods in native C++ to create advanced 3D audio effects [15], [19], [20], including coverage of the front, back, and sides of the listener. Recent versions of the library supports HRTF (Head Related Transfer Funtion), based on the KEMAR HRTF dataset by MIT [21] (only available when using 44100Hz playback), which provides much more versatility in the perceived placement of sounds, giving the listener the impression they are also locate above and below his/her head.

Although this solution can represent a 360°variation of the sound's azimuth quite realistically (including behind the user),

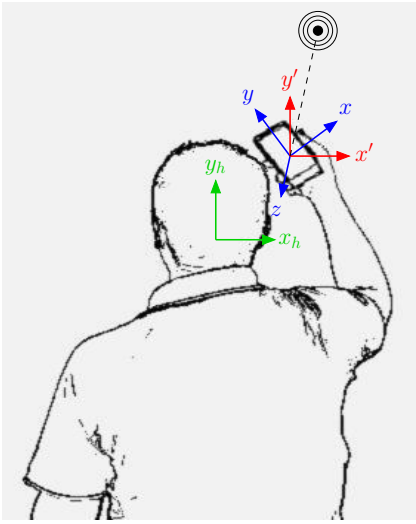


Fig. 4. Smartphone's frame of reference $\langle x, y, z\rangle$; 3D sound frame of reference $\langle x', y', z\rangle$; and head frame of reference $\langle x_h, y_h\rangle$, with $x_h$ oriented from the left to the right ear. The concentric circles on the top-right indicate the sound source (i.e. visual target).

the performance on the vertical plane is usually not as good, making it difficult sometimes to distinguish between sounds coming from the bottom or the top. Besides technological limitations, research on 3D sound localization suggests the problem is caused also by the difficulty of humans to estimate the direction of vertical sound variations [16]. A possible solution could be the use of difference sound effects, depending on the current zenith $\phi_e$. However, for sake of simplicity, and motivated again by the desire to keep the information for the user at a minimum, the same sound (a simple *click* in the current implementation[1]) was used, independently of the source direction.

### B. Frame of Reference

An important choice for the successful application of the 3D audio interface is the frame of reference adopted for the sound spatialization. Recall from Section II that a sound source corresponds to a particular visual target the smartphone's camera should be pointed to, and that a hand-held devices is here considered. Because of the 6 DoF of the camera, the visual target of interest (e.g. image feature, object, etc.) may be observed from different orientations and distances. Dealing with rotation and scale variations is usually not a problem for modern computer vision algorithms. However, rotations of the smartphone about the $z$ axis (i.e. normal to the smartphone's display, see Fig. 4) can be problematic if one has to make sense of the sounds in this frame of reference.

When a user observes the world through the "virtual" eye of a camera, indeed, he/she can usually project him/herself quite easily into the respective frame of reference, without feeling disoriented (at least for moderate movements of the camera or the user). However, the same is not true for sound perception: people do not usually tilt their heads left or right to have a better listen of the sound (although they might do so if they use one ear only). If the source is mapped into the device frame of reference, as is usually the case for head-mounted solutions in the literature, and the smartphone is pointed in such a way that the same orientation would be very unnatural for a human head, then it is difficult for the user to interpret the audio signals correctly.

Let's explain this with an example: if the sound source in Fig. 4 is located on the top of the human head and the smartphone's camera is directed towards it, but the $y$ axis points left (i.e. the smartphone is held so the angle about $z$ is ~45°or more), then the user will have the impression that the sound source is on his/her right, rather than above him/her.

To avoid this problem, in the current system the sound sources are mapped, through an opportune coordinate transformation, into a zero-roll frame of reference, i.e. one in which the $x'$ axis is always parallel to the ground. In this way, the 3D audio signals will only suggest azimuth and zenith of the visual target as if the user's head was virtually located in place of the smartphone, but without being tilted left or right.

## V. Conclusion

The paper introduced the solutions adopted for a prototypical mobile assistive device to aid the indoor navigation of VI people. It presented in particular a possible multimodal interface for an active vision system with human-in-the-loop, implemented as a smartphone application. The system differs from previous vision-based solutions in that it guides the user towards interesting visual features, rather than relying on image acquired by fixed wearable cameras. The flexibility introduced by such a system allows for a better coverage and detection of features around the user, either for localization or object identification, in particular when the latter are very close to the blind person. The proposed interface is strongly based on 3D audio spatialization, which provides a feasible and economically convenient alternative to other wearable and haptic devices. It has currently reached the evaluation stage, in which experiments with several blind-folded users will be carried out to measures performances in terms of accuracy, success rate (e.g. visual targets correctly localized), user response time, information trade-off (e.g. maximum number of detectable targets per time unit), etc.

The proposed solution is specifically designed to work in combination with a visual-inertial localization system (e.g. [22], [23]) and with an object identification algorithm for obstacle and hazard detection, in order to assist VI walking indoor. It is also thought as a means to facilitate the implementation of perception algorithms to solve the "last 10 meters" problem, helping VI people to locate objects and places of interests in their proximity (e.g. doors, vending machines, staircases, etc.). Steps in this direction are being taken for extensions of the system in the near future. In the long term, automatic user adaptation and self-regulation of the cognitive load are other interesting research areas to be investigated.

[1]Different sounds and musical tones could be otherwise used, but the final choice depends on the preference of VI users, as already highlighted in [16].

## REFERENCES


[1] R. Bajcsy, "Active perception," *Proc. of the IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[2] E. Rivlin and H. Rotstein, "Control of a camera for active vision: Foveal vision, smooth tracking and saccade," *International Journal of Computer Vision*, vol. 39, pp. 8–96, 2000.

[3] R. Manduchi and J. Coughlan, "(Computer) Vision Without Sight," *Communications of the ACM*, vol. 55, no. 1, pp. 96–104, 2012.

[4] L. Ran, S. Helal, and S. Moore, "Drishti: An integrated indoor/outdoor blind navigation system and service," in *Proc. of the 2nd IEEE Annual Conf. on Pervasive Computing and Communications (PERCOM)*, 2004, pp. 23–30.

[5] J. Martinez and F. Ruiz, "Stereo-based Aerial Obstacle Detection for the Visually Impaired," in *Workshop on Computer Vision Applications for the Visually Impaired*, Marseille, France, 2008.

[6] A. Davison, W. Mayol, and D. Murray, "Real-time localization and mapping with wearable active vision," in *Proc. of the 2nd IEEE and ACM Int. Symposium on Mixed and Augmented Reality*, 2003, pp. 18–27.

[7] J. Coughlan, R. Manduchi, and H. Shen, "Cell phone-based wayfinding for the visually impaired," in *Proc. of the 1st Int. Workshop on Mobile Vision*, Graz, Austria, 2006.

[8] I. Apostolopoulos, N. Fallah, E. Folmer, and K. Bekris, "Integrated online localization and navigation for people with visual impairments using smart phones," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2012, pp. 1322–1329.

[9] R. Manduchi, S. Kurniawan, and H. Bagherinia, "Blind guidance using mobile computer vision: a usability study," in *Proc. of the 12th Int. ACM SIGACCESS Conf. on Computers and Accessibility*, New York, NY, USA, 2010, pp. 241–242.

[10] V. Pradeep, G. Medioni, and J. Weiland, "Robot vision for the visually impaired," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 15–22.

[11] J. Ward and P. Meijer, "Visual experiences in the blind induced by an auditory sensory substitution device," *Consciousness and Cognition*, vol. 19, no. 1, pp. 492–500, 2010.

[12] G. Bologna, B. Deville, and T. Pun, "On the use of the auditory pathway to represent image scenes in real-time," *Neurocomputing*, vol. 72, no. 4-6, pp. 839–849, 2009.

[13] B. Katz, P. Truillet, S. Thorpe, and C. Jouffrais, "NAVIG: Navigation Assisted by Artificial Vision and GNSS," in *Pervasive Conf.- Workshop on Multimodal Location Based Techniques for Extreme Navigation*, Helsinki, FI, 2010.

[14] A. Davison and D. Murray, "Simultaneous localization and map-building using active vision," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 865–880, 2002.

[15] G. Hiebert, "OpenAL 1.1 specification and reference," Creative Labs, Tech. Rep., 2005. [Online]. Available: http://connect.creativelabs.com/openal/Documentation

[16] M. Bujacz, M. Pec, P. Skulimowski, P. Strumillo, and A. Materka, "Sonification of 3d scenes in an electronic travel aid for the blind," in *Advances in Sound Localization*, P. Strumillo, Ed. InTech, 2011, ch. 14, pp. 251–268.

[17] Android SDK. (Accessed on 20th May 2013). [Online]. Available: http://developer.android.com/sdk

[18] Ivona text-to-speech. (Accessed on 20th May 2013). [Online]. Available: http://www.ivona.com/

[19] OpenAL Soft. (Accessed on 20th May 2013). [Online]. Available: http://kcat.strangesoft.net/openal.html

[20] OpenAL Soft with Android support. (Accessed on 20th May 2013). [Online]. Available: https://github.com/AerialX/openal-soft-android

[21] HRTF Measurements of a KEMAR Dummy-Head Microphone. (Accessed on 20th May 2013). [Online]. Available: http://sound.media.mit.edu/resources/KEMAR.html

[22] M. Li and A. Mourikis, "Vision-aided inertial navigation for resource-constrained systems," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Oct. 2012, pp. 1057–1063.

[23] L. Porzi, E. Ricci, T. Ciarfuglia, and M. Zanin, "Visual-inertial tracking on Android for Augmented Reality applications," in *IEEE Workshop on Environmental Energy and Structural Monitoring Systems (EESMS)*, Sept. 2012, pp. 35–41.