Low-Latency Human-Computer Auditory Interface Based on Real-Time Vision Analysis

Florian Scalvini, Camille Bordeau, Maxime Ambard, Cyrille Migniot, Julien Dubois

LOW-LATENCY HUMAN-COMPUTER AUDITORY INTERFACE BASED ON REAL-TIME VISION ANALYSIS

Florian Scalvini[1]*, Camille Bordeau*[2]*, Maxime Ambard*[2]*, Cyrille Migniot*[1]*, and Julien Dubois*[1]

[1]ImViA EA 7535 - Univ. Bourgogne Franche-Comté, Dijon, France
[2]LEAD CNRS UMR 5022, Univ. Bourgogne Franche-Comté, Dijon, France

ABSTRACT

This paper proposes a visuo-auditory substitution method to assist visually impaired people in scene understanding. Our approach focuses on person localisation in the user's vicinity in order to ease urban walking. Since a real-time and low-latency is required in this context for user's security, we propose an embedded system. The processing is based on a lightweight convolutional neural network to perform an efficient 2D person localisation. This measurement is enhanced with the corresponding person depth information, and is then transcribed into a stereophonic signal via a head-related transfer function. A GPU-based implementation is presented that enables a real-time processing to be reached at 23 frames/s on a 640x480 video stream. We show with an experiment that this method allows for a real-time accurate audio-based localization.

Index Terms— Auditory sensory substitution, people detection, wearable assistive device, real-time processing

1. INTRODUCTION

A recent study estimates that, despite advances in preventive treatments, the growth and aging of the population should lead to an increase from 43 millions in 2020 to 61 millions of blind people in 2050 [**?**].

For this people the lack of visual information leads to multiple challenges and daily tasks such as walking in a non-familiar environment without colliding an obstacle remains challenging. The white cane and the trained dog are the classic methods used to remedy this problem. Although these means are very widespread and effective for moving in an unfamiliar environment, they do not provide the user with all the useful information to know his environment. The white cane limits the perception to close objects and the trained dog requires a long training and is relatively expensive.

Assistive technology systems have been the subject of research since decades. Recent advances in image processing methods and the performance of embedded modules in terms of computing power, consumption and miniaturization now offer new possibilities. Among these devices, sensory substitution systems (or SSDs), is a category that uses the brain ability to construct a representation of the world based on a new sensory encoding. These SSDs convert information normally acquired through vision into a signal designed for another sensory modality, mainly auditory or tactile.

For visuo-auditory SSD, this process is called sonification [1]. In the vOICe [2], the pioneer system of visual sensory substitution by sonification, the pixels of a 2D image are sonified according to their positions and luminance.
New sonification protocols provide a stereophonic sound that gives the ability to locate a static or moving object in 2D or 3D scene [3] or to move [4]. The spatial position encoding into a stereophonic sound is computed by Head-Related Transfer Functions simulating the reflections of the sound on the human body before entering the two cochlea. Vision-based and sound generation processing in real-time induce significant latency that should be minimized. For instance setups of some studies [5, 6] acquire data from a smartphone and perform the processing on a laptop in a backpack . Deporting the computing unit to a remote server reduces weight and space requirements and increases autonomy. However, a remote sensing system causes transmission delays and requires a constant connection to operate.

Human-computer auditory interfaces have been designed to localise person [5] based on artificial vision. These systems enable a single class, the obstacles, to be localised in an 2D environment. Alternative methods [7] demonstrate that semantic information could be used in such interfaces as well to perform the environment's perception. The resulting systems enable each significant word to be replaced by discriminant sounds (for instance using different musical instruments). A specific short-sound dictionary has been designed in order to preserve partially the richness of a verbal language meanwhile with respect of the human's physiological reaction-time. Based on this approach, a recent study [1] proposes a SSD using a similar sonification protocol where each object (car, people,...) is associated to a short one-second sound. Nevertheless, a powerful PC platform does not allow real-time to be performed as the algorithm's complexity requires high computational resources. Despite high

Thanks to the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) which are supporting financially this research.

performances in terms of people localisation, the mobility and usability of such a system is then reduced.

In this paper, we propose a human-computer auditory interface system which enables 3D localisation of person based on a RGB-D camera. We focus specifically on person localization since it is one of the most frequently encountered situation during urban walking. Contrary to state-of-art other approaches, the global processing is performed in real-time on standard computational platforms as well as on processing units dedicated to embedded system designs. Hence, a new generation of human-computer auditory interface design is clearly targeted to ease the daily life utilisation and to proposed a mobile and compact system with lower-power consumption, and therefore, higher autonomy. Considering the problems related to a remote transmission, we integrate all video and sound processing on the embedded system.

2. METHOD

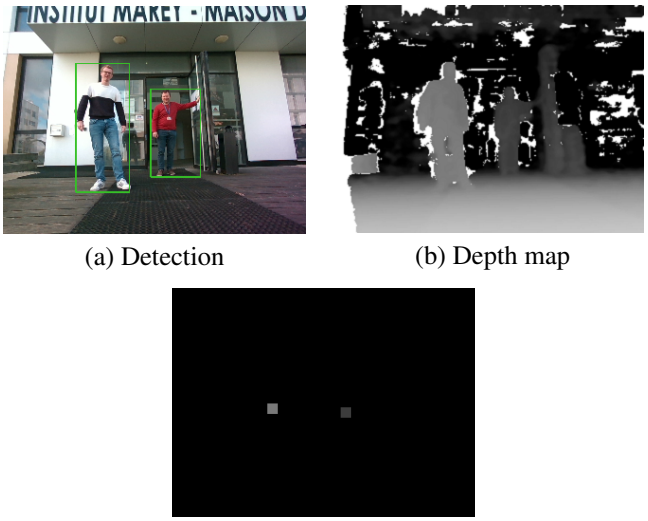
In this section, we describe our method for localizing person in a 3D environment and generating the associated stereophonic sound. Two pipelined processing stages are performed to extract the 3D person localisation.

At the first processing stage: a trained convolutional detection network (CNN) estimates the positions of the persons from the 2D scene. We chose a CNN model based on the architecture You Only Look Once (YOLO) because it has the best framerate [8] for a detection network that produces robust detection on multiple scales. A bounding box detected around the visible part of the person is available as the output of this first processing stage (Figure 1.a). The bounding box size, the 2D positions of the corresponding centroid, and the confidence score are then available for each detection.

The second image processing stage consists in the distance-to-user estimation using the corresponding depth map. This information is provided by the stereoscopic camera (Figure 1.b) which is associated to the rgb standard imaging system.

Finally the position of the centroid is extracted ((Figure 1.c) and sonified with a stereophonic signal generated according to the 3D people localisation. Based on the sound spatialization, the user is then able to localise the targeted person in the 3D space.

Our method of sonification is based on the LibreAudioView system [3] in which each visual object generates an audio signal. In this study, the sonification of the localisation of the target was done as follows: For each video frame the pixel corresponding to the center of the bounding box was extracted if a person has been detected. Given the visual field of the camera, the coordinate of the pixel is mapped on spherical coordinates on a sphere of 2 meter radius centered at the camera. We then used the two meters HRIR data set recorded in an anechoic chamber [9] to spatialize a brief (33ms) monophonic 440 Hz sound with a 5ms cosine fade-in and fade out. For each possible pixel position, we pre-calculated its HRIR spatialization based on the Input responses of the corresponding azimuthal position in the HRIR data set. The amplitude of sound is modulated depending on the distance which separate the target from the user using an inverse square law $A = 1/d^2$ where A is the amplitude and d the distance between the user and the target.



(a) Detection (b) Depth map

(c) Pixels to sonify

Fig. 1. Overview of the method: first a CNN estimates the positions of the person from the color image (a); then the distance-to-user is extracted from the depth map (b) to generate the pixels to sonify (c).

3. REAL-TIME IMPLEMENTATIONS

For our system, we have implemented a CNN trained on the Microsoft Coco database [10]. This database labels 80 different object classes on the 328124 available images. There are 250000 occurrences of person among the 1.5 million of detected objects. Yolov5-small CNN has been selected for implementation in order to propose in fine an embedded system and respect the application constraints. The CNN has been trained on 640×640 resolution images.
The video acquisition is performed using an Intel RealSense D455 stereoscopic camera. Efficient acquisitions can be obtained, equally in outdoor or indoor environments, with a Field of View (FOV) of 87° × 58° and an ideal range of 0.6 to 6m. Both color and depth image were synchronously captured with a resolution of 640×480 pixels at 30 frames per second. The color image is resized to respect the training resolution of the CNN.

As previously mentioned, low-latency system is required in regards to the application constraints. Therefore we have proposed an optimization of the software solution. More

precisely, we aim to minimize the delay between an image acquisition and the sound transmission to the user. The LibreAudioView sonification architecture has been optimized in a previous paper [11]: the optimization of the sonification software has reduced the required processing time by 86% compared to the original version [3]. After the optimization of the sound generation stage, the image processing part represents 95% of the global processing time on a standard PC platform (Intel Core i7-6700HQ processor : 4 Cores - 8 Threads, 2.60 GHz; 16 GB RAM). Therefore, we propose GPU-based implementations to reduce the processing speed. Indeed, the specific multi-core GPU architecture is particularly adapted to regular tasks and to the intrinsic parallelism of the selected algorithm. Finally, we propose a second implementation based on a GPU target that is dedicated to embedded system designs. The goal is to demonstrate that low-latency solution can be designed around such a target to propose a compacted and embedded system. Moreover, the power consumption has been adjusted to increase the system's energy autonomy respecting the application's performance constraints. Different optimisations are then proposed, as the adaptation of the data dynamic, to decrease the system's latency.

First, a comparison is proposed between a standard CPU-based implementation and standard GPU one. As previously, the CPU implementation is based on an Intel Core i7-6700HQ processor (4 Cores - 8 Threads, 2.60 GHz; 16 GB RAM); Meanwhile the GPU implementation is based on a Nvidia GTX 1070 GPU (2048 CUDA Cores, 6.738 Tflops, 8GB VRAM). A laptop is integrating the two targets. The neural network is implemented on both CPU and GPU targets with the Libtorch library (C++ version of Python). The two-first columns of the Table 1 represent the comparison between the two targets and summarizes the performance of the YOLOv5-small and its impact on the overall operation of the sonification device. Please note that the comparison is realized with sequences of images to avoid that the camera frame rate limits the measurement. Considering these results, the inference time of the YOLOv5-small network on a standard computer processor does not enable real-time performances with 640×640 images to be reached whereas on a GPU target this can be achieved. Moreover, the inference time of the CNN on GPU has been reduced by converting the model with the TensorRT SDK. TensorRT is an inference optimizer on Nvidia platforms. A significant gain of 53% between the optimized and non-optimized model is obtained as depicted in the third column of Table 1.

Considering the problems related to a remote transmission, we favour the development of an autonomous device. The low-latency processing is a keystone to reach system's autonomy and hence providing system user's security. A GPU-based implementation represents an appropriate solution to accelerate the processing and therefore a pertinent solution to develop in-fine a wearable device. Indeed, some GPU targets are dedicated to embedded system design by offering a high trade-off between high processing performances and power consumption. Hence, an Nvidia Jetson TX2 (8GB RAM, 256 CUDA Cores, 1.33 TFlops) embedded module has been used. Moreover such target supports and benefits from TensorRT optimization on CNNs. The number of Flops on the embedded card is 5 times lower than on the previous Nvidia GTX 1070 GPU board nevertheless other optimizations are available. Indeed, Nvidia proposes through its Jetson modules and its latest graphics cards (RTX series), the possibility to modify the dynamic range of the CNN's weights. It can be fixed to 16-bits floating format (FP16 : Half precision) instead of 32-bits (FP32). The FP16 configuration decreases significantly the inference time of the CNNs and the memory requirements. On the COCO 2017 validation database, both the configuration provide an accuracy of 55.4% for an overlapping of 50%. Moreover, the selected embedded module, running on Jetpack 4.6 (Ubuntu 18.04, Cuda 10.2, TensorRT 8.0.1), can operate in two power modes: Max-Q of 7.5W (5.5V) and Max-P of 15W. The system's energy autonomy is estimated with a 10 000 mAh - 12 Volts commercial battery. The operating life of the system at full power (TX2 module: 15 W & Realsense camera: 2.335W, the consumption of headphones is ignored) is estimated at 6 hours 55 minutes. The estimated battery life at low power (7.5W + 2.335W) is 12 hours 11 minutes.

	CPU	GPU	
		LibTorch	TensorRT
Yolov5-small (ms)	135	16.3	7.5
Global processing (ms)	142	23.6	15.3

Table 1. Processing time of the YOLOv5-small using a laptop with the overall system (CNN implemented on three targets).

The Table 2 summarizes the impact of the CNN on the embedded target using the two different power modes and considering the two proposed dynamic ranges. An inference of the YOLOv5-small network is lower to the camera framerate on a Nvidia TX2 with a resolution of 640×640 in input.

However the experimental system, with half-precision and maximum power, provides an audio perception equivalent to real-time. The use of this low-power mode generates a larger latency but still enables performances compatible with the application's constrains to be obtained. Hence, this choice of mode is pertinent considering the significant gain in term of operating life.

4. EXPERIMENT

We measured the capabilities offered by such an auditory sensory substitution device in a task consisting in localising human bodies that are in close vicinity. Based on the spatial

	FP32 15W	FP32 7.5W	FP16 15W	FP16 7.5W
Yolov5-small (ms)	55	71	35	47
Global Processing (ms)	62	80	43	56

Table 2. Processing time of the YOLOv5-small on the embedded target in comparison with the overall system. Two power modes and two dynamic ranges are proposed.

information extracted from the 3D video stream, the position of a standing person was transmitted using the spatialized audio encoding described in the section 2.

For this purpose, we used an experimental setup based on an HTC Vive system to track the position of the sensory substitution user's head during a localisation task under two conditions. In the auditory condition, ten blindfolded participants sitting on a 360° revolving chair were placed at the center of a 4m×4m area. A sensory substitution system and a HTC position tracker were fixed on the participant's forehead. The target to localise was a person wearing a second HTC tracker on his sternum, standing at random places 2 meters away from the revolving axis of the chair, on 8 equally spaced positions with a fixed angle gap of 45° as presented in the figure 2.

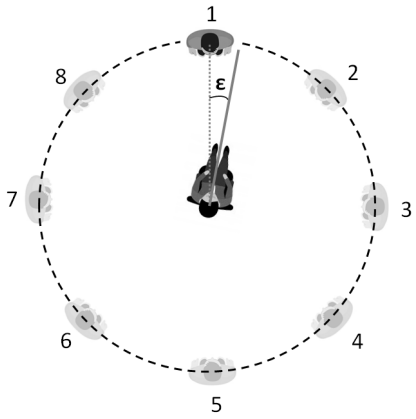


Fig. 2. Experimental setup used to measure the localization abilities. The participant sits at the center of the experiment area on a revolving chair. A standing person randomly changes its standing position among 8 marked places [1...8] equally spaced on a two meters radius circle. We measured the azimuth angle error ε.

Each trial was first composed of 10 seconds of white noise loud enough to cover the sound that might be produced by the target person changing its standing position. After these ten seconds the participant had to rotate on the chair in order to find the target and place it in front of him based solely on the auditory indications provided by the substitution system. The validation was given by pressing on a joystick button. Two trials were performed for each position. In the visual condition, the 10 participants were not blindfolded and the same task was reproduced only with vision, i.e. pointing the head towards the standing person solely using visual feedback.

Results presented in the figure 3 show mean azimuth angular error for each target position in the auditory condition. In this condition, mean azimuth angular error was 6.72° ± 5.82. As expected, mean azimuth angular error in the visual condition was approximately 2 times smaller (2.85° ± 1.99). Despite this difference, these results show that participants could localise a person with high accuracy using our auditory sensory substitution device.

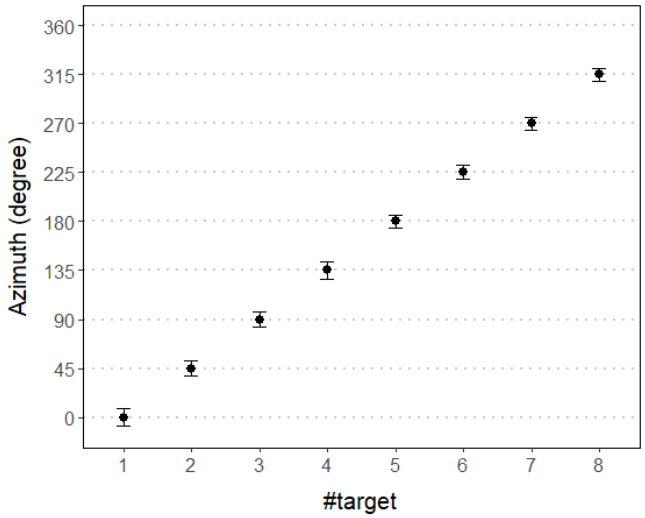


Fig. 3. Azimuth of each target (black dot) with the associated mean value of the azimuth angular error (vertical bar) in the auditory condition.

5. CONCLUSION

In the context of visually impaired people assistance, there are strong constraints in terms of latency, autonomy and portability of the system. In this paper we introduced a new system where the 3D position of person detected by a CNN is sonified into a stereophonic sound. First, tests on laptop have shown that real-time performances with 640×640 images have been achieved on a GPU target. In a wearable device, our system provides an audio perception equivalent or close to real-time. Two power modes and two dynamic ranges allow a compromise between latency and operating life to be adjusted according to the user's preferences. Finally the capacity of a user wearing our device to perceive a person's position has been evaluated and demonstrated experimentally. Future work will extend this protocol to new classes to sonify and enrich the audio signal by a verbal expression of specific events.

6. REFERENCES

[1] Angela Constantinescu, Karin Müller, Monica Haurilet, Vanessa Petrausch, and Rainer Stiefelhagen, "Bring the Environment to Life: A Sonification Module for People with Visual Impairments to Improve Situation Awareness," in *International Conference on Multimodal Interaction*. 2020, pp. 50–59, ACM.

[2] Peter B.L. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, 1992.

[3] Maxime Ambard, Yannick Benezeth, and Philippe Pfister, "Mobile Video-to-Audio Transducer and Motion Detection for Sensory Substitution," *Frontiers in information and communication technologies*, vol. 2, 2015.

[4] Barthélémy Durette, Nicolas Louveton, David Alleysson, and Jeanny Hérault, "Visuo-auditory sensory substitution for mobility assistance: testing TheVIBE," *Workshop on Computer Vision Applicationsfor the Visually Impaired*, pp. 1–13, 2008.

[5] Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia, "DEEP-SEE: Joint Object Detection, Tracking and Recognition with Application to Visually Impaired Navigational Assistance," *Sensors*, vol. 17, no. 11, pp. 2473 1–24, Oct. 2017.

[6] Matteo Poggi and Stefano Mattoccia, "A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning," in *IEEE Symposium on Computers and Communication*, Messina, Italy, 2016, pp. 208–213.

[7] Guido Bologna, Benoît Deville, Thierry Pun, and Michel Vinckenbosch, "Transforming 3D Coloured Pixels into Musical Instrument Notes for Vision Substitution Applications," *EURASIP Journal on Image and Video Processing*, vol. 2007, pp. 1–14, 2007.

[8] Andrei-Alexandru Tulbure and Eva-Henrietta Dulf, "A review on modern defect detection models using DCNNs – Deep convolutional neural networks," *Journal of Advanced Research*, 2021.

[9] Hagen Wierstorf, Matthias Geier, and Sascha Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," *Journal of the audio engineering society*, 2011.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár, "Microsoft COCO: Common Objects in Context," *European Conference on Computer Vision*, pp. 740–755, 2015.

[11] Maxime Ambard, "Software Design for Low-Latency Visuo-Auditory Sensory Substitution on Mobile Devices," *Computer and Information Science*, vol. 10, no. 2, pp. 1, 2017.