MODAL—a Multi-modal Sensing Platform for IoT Applications

N. Cantale, P. Nussbaum, P. Molticek

Robust sensors for people tracking and health state monitoring are needed to allow the ill and elderly to maintain their independence and to relieve the burden on the healthcare system. This work, in a collaboration with Idiap, is focused on merging the output of several sensor types to provide accurate monitoring and alarm raising in the context of eldercare.

CSEM has a track record of developments in building occupancy detection and monitoring using embedded vision. CSEM and Idiap's recent collaborations with industry has revealed an interest and potential of autonomous smart sound devices. Applications exist in security, surveillance and eldercare. Given the aging population, the segment of elder care and more specifically key technologies that can detect and manage critical situations (falls, distress, …) is expected to grow.

Visual detection and localization of people in buildings finds limitations in many ambiguous situations where the absence of motion or presence of artefacts are reducing the accuracy and robustness of detection. Furthermore, the information necessary to evaluate the seriousness of a situation (people care) is often absent from the images alone. The goal of the present project is to bring complementary information from the analysis of ambient sound and combine it with the visual analysis. Going multi-modal improves substantially the robustness of the system and provides the necessary features (speech analysis, sound localisation, and automatic speaker and speech recognition) to address several market segments.

The combination of the visual and sound information is done on an embedded platform based either on the Nvidia TX2 or Nano IoT solutions. The platform can host up to 3 MIPI cameras or several USB cameras. The audio comes from a 6 channels microphone array from MiniDSP and the communication is done through Ethernet or a Wifi access point. The platform is designed to easily host additional sensors in the future, in order to improve their input for better accuracy or functionality (*e.g.*, RF sensing).

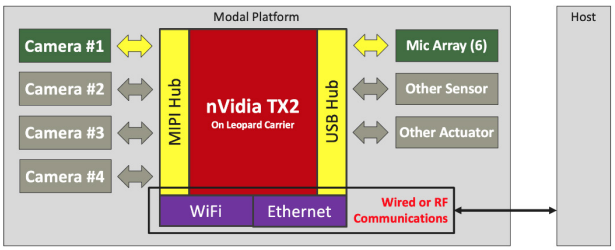

Figure 1: The Modal platform.

Figure 2: The housing for the Modal platform.

The platform is mounted inside a 100 x 100 x 70 mm housing, containing the Nvidia module, the carrier board, the camera, the microphone array, a USB hub, a Wifi antenna. There is extra-space for optional modules like a LoRa modem.

Two machine learning algorithms run side-by-side on the platform. The first one does the people detection, based on a network previously developed by E. Türetken [1]. The system achieves a detection precision of 95% and a recall of 94% on the test dataset. It can distinguish people with a spatial resolution of less than half a meter on the floor, sufficient for most applications. The trained model has 37 thousand weights, which takes less than 200 KB memory to store, and can run rea-time with limited computational resources.

The information from the visual monitoring is sent to the second algorithm, which analyses the audio inputs and does the following tasks:

- Audio activity detection detects the presence of active speakers and discriminates from other sound sources;
- Speaker identification: identifies the speaker from a list of registered users;
- Keyword spotting: detects commands from a list of keywords;
- Source localization: performs the 3D location estimation of the speaker using beamforming.

The output is then an audio command from an identified speaker, with its location, which can be sent to a server or another device.

Future works include further miniaturization of the platform and improved algorithms natively implementing tracking capabilities.

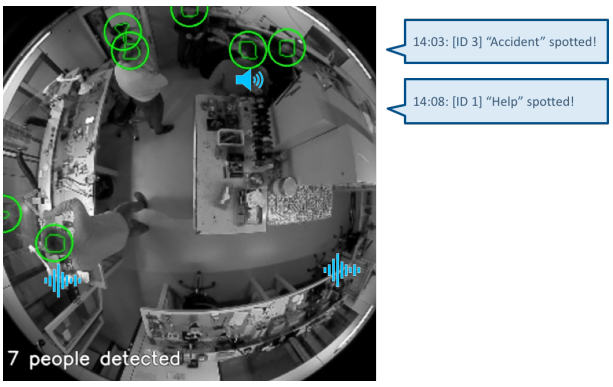

Figure 3: Illustration of the audio and video fusion. Image from the MODAL platform. The people are detected (in green), the speaking (blue speaker icon) and non-speaking (blue wave icon) sound sources are separated, the speakers are identified, and the keywords are spotted (bullets on the right).

[1] E. Türetken, L. A. Dunbar, "Efficient Deep Learning Algorithm for Person Detection from Ceiling-mount Cameras", CSEM Scientific and Technical Report (2018) 99.