Realisation of an Audio & Video Laboratory for Precise Object Localisation and Tracking

Robert Manthey[1], *Hussein Hussein*[2], *René Erler*[1], *Richard Siegel*[1], *and Danny Kowerko*[1]

[1]*Department of Computer Science, Chemnitz University of Technology, Chemnitz, Germany*
[2]*Department of Literary Studies, Free University of Berlin, D-14195 Berlin, Germany*
firstname.lastname@informatik.tu-chemnitz.de, hussein@zedat.fu-berlin.de

Abstract: This paper presents the realisation of an audiovisual laboratory for detection, localisation, classification and tracking of objects in indoor environments using visual as well as audio information. The laboratory is property of the endowed junior professorship Media Computing at the Chemnitz University of Technology. Visual information is retrieved by 10 optical embedded smart stereo sensors (Intenta S2000). The visual sensors can be used to trigger an identification process of objects and persons as well as a tracking operation to follow them in the field of view. For audio signal processing, a total of 64 microphones and 16 loudspeakers are used. Three microphone arrays are constructed using three different types of microphones. The camera sensors and microphones can be mounted in different positions, directions and heights. The loudspeakers can be set up freely within the tracking area using standard monitor stands. Within a separate air-conditioned server room, all audio signals are pre-amplified and AD/DA converted using industry standard rack-mounted audio hardware. A server cluster and workstations with high-end Nvidia P6000 graphics cards provide the raw processing power for all processing tasks within the scope of use cases for the laboratory. Software which is available for users in the laboratory comprises commercial products, such as Steinberg Cubase 8.5, as well as self-developed solutions, such as an audio & video localisation and annotation tool. This will be complemented by documenting acoustic sources (e.g. loudspeakers) and detectors (microphones) into a Blender-animated 3D laboratory using the video sensor data. Interested individuals will also have the opportunity to make use of the laboratory's extra peripheral equipment, e.g. a Yamaha DGX-660 Keyboard, further types of mobile wireless microphones, HTC Wive VR glasses and wireless Internet of Things (IoT) smart home sensor probes for motion capturing. The active tracking area comprises a total of 82 sensors.

1 Introduction

Visual, acoustic as well as combined audiovisual information can be used to identify objects and to estimate their positions in space. Today, many devices and applications such as mobile robots, surveillance systems and human-computer interaction benefit from object localisation and tracking. Microphones in specially designed configurations receive acoustic signals and form the basis for object localisation by evaluating propagation delay between sound source and different microphones. Visual localisation can be processed similarly. Cameras capture the light within their field of view and subsequently image processing methods are used to perform the task of desired object identification and tracking. In order to be able to reproduce real world events, a laboratory is designed and set up to contain a system of passive sensors. Measurements made in such a controlled environment favour evolution, extension and monitoring of algorithms approximating the real world. A common example for multi-sensor applications using microphones and cameras is the smart home or ambient assisted living.

Other existing facilities which rank within the scope of applications of our laboratory are the Fraunhofer IDTM Sound-Technologies-Lab [1] and the Cognitive Systems Lab at BTU Cottbus-Senftenberg [2]. The Fraunhofer IDTM Sound-Technologies-Lab is mainly focused on audio systems and offers e.g. a showroom for ambient assisted living solutions, auditory booths, an anechoic chamber, a reverberation room and a wind tunnel. The BTU Cottbus-Senftenberg Cognitive Systems Lab is specialised for development of intuitively controlled, adaptive self-learning systems and is equipped with audio as well as video technology. It consists of a specially acoustically treated room, where reverberation time can be variably adjusted by detachable absorber panels. Within this room an additional soundproof cabin is situated, which can be used as a recording or technical room.

The design of our laboratory for object and person localisation and tracking using acoustic and visual information is presented in [3], describing the main components (64 microphones, 16 loudspeaker, multi-channel interface, 8 audio interfaces, and audio clock generator) and the connection between them to transmit the audio data from microphones into the computer and from the computer to the loudspeakers. The design of the video system is presented schematically and will be reviewed in more detail within this article.

The preliminary works within the project for object localisation and tracking as well as the analysis of object behaviour are: acoustic event classification for utilisation in the sector of ambient assisted living or general indoor scenarios [4, 5], the effect of microphone array geometry in algorithms for acoustic source localisation [6], and classification of bird sounds using convolutional neural networks (CNN) [7].

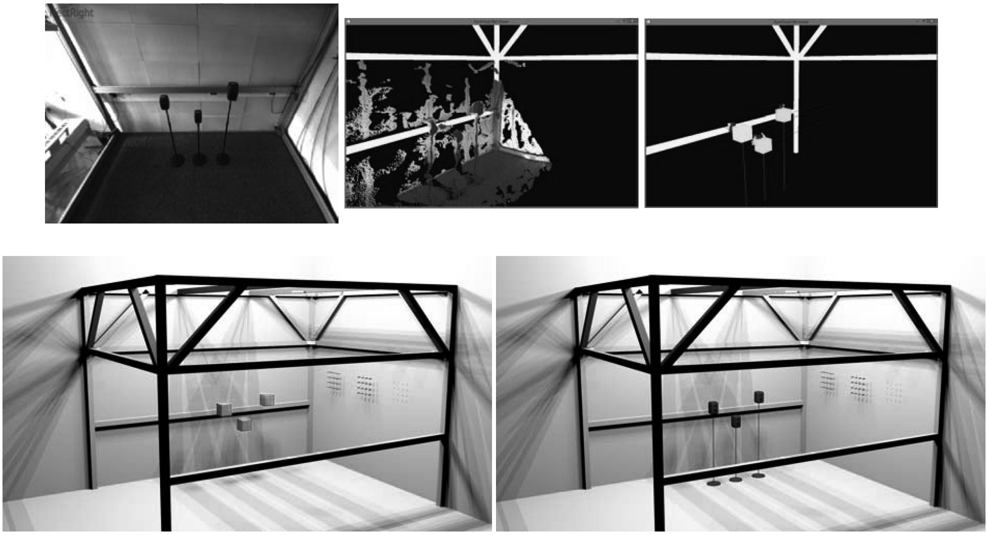

Figure 1 – Example of 3D localisation workflow from top left to bottom right: static objects like speakers localised using point cloud information (for details, see text)

2 Audio & Video Laboratory

This section describes the technical video and audio equipment of the laboratory in the context of its fields of applications, which are mainly the localisation and tracking of objects and persons in indoor environments. Further, we combine the smart localisation information from the visual stereo camera sensors itself and state of the art video analysis using CNN based algorithms for object classification and localisation within our video footage. The laboratory consists of two rooms. The first one (L × W × H = 7.2 × 6 × 4 m) with an active tracking area of (L × W × H = 4 × 3.5 × 3.5 m) is used for the installation of acoustic and optical sensors. This room is equipped with sound insulation materials to dampen the acoustic reflections. The second room is a cooled server room for mounting of the audio interfaces.

2.1 Video System

Visual information is retrieved by recording videos in HD resolution from 10 optical smart embedded stereo sensors (S2000, Intenta GmbH) with the properties shown in Table 1. The information produced by each sensor contains the video streams of the two camera lenses, but also the distance of all detected foreground objects as well as classifications and coordinates of bounding boxes of recognised objects. Note that different background models need to or may be applied to exclude static objects which are interpreted as background. The classification can distinguish person behaviour in lying, sitting and standing. Data is transmitted via Ethernet to preprocessing systems to combine the different views, reduce the amount of data using FFmpeg compression and to create a modelled 3D representation of the scene, e.g. in Blender [8] as illustrated in the bottom of Figure 1 and the left of Figure 3. Point cloud sensors are utilised in order to retrieve the 3D positions of objects placed in the laboratory. The raw point cloud data is split into subsections created through an evenly distributed three dimensional grid. In order to be able to focus on the objects which will be placed in the laboratory, a mask for the point cloud data is created first. The mask is a list of all the subsections which hold points when an empty laboratory is scanned. After applying the mask and thus removing the points representing the laboratory itself, only objects which were not part of the empty laboratory will be visible in the point cloud. To remove noise and insignificant parts of the remaining point cloud a filter is implemented by removing every subsection of the cloud which holds less then a specified number of cloud points. This filter value can be adjusted to achieve the desired result. At last, object markers are placed based on the location of the remaining subsections. In order to be able to utilise such found object positions, the raw point cloud should first be synchronised with the laboratory model. Although the measurement precision of this method is still being improved, 3D models of laboratory settings can be exported using the OBJ-format. Each object position relative to the laboratory is also available. The workflow is exemplified for speaker localisation in Figure 1 from top left to bottom right, where i. shows the rectified real image, ii. a 3D model of the laboratory cage with an aligned point cloud, iii. remaining point cloud after applying a mask with cubic object markers iv. Blender imported cube object markers including the laboratory cage and v. 3D model with previously modelled speakers placed centrally at the cube locations. The 3D scene helps planning experiments, e.g. to control occultation of persons and objects by obstacles and can increase the quality of tracking of persons. Subsequent 3D analyses of the preprocessed data of the scene facilitates better understanding of the activity of persons and their behaviour as function of space and time. Recording and (pre-)processing up to 10 stereo camera videos entails a vast amount of data which are transferred to our storage and analysis cluster. The analysis of the scene, the person activity and their interaction will be processed by the cluster and/or GPU workstations.

Documentation of activities in 3D supports the reproduction and interpretation of activities and defines our ground truth for evaluations. With our virtual, Blender-based model all activities and combinations can be realised to produce synthetic data for learning, evaluation and improving algorithm quality.

Table 1 – Specification of optical smart stereo sensors

Weight	500 g
Unit Size	200 × 70 × 33 mm
Field of view	97°
Area of detection	2 × 2 m to 4.5 × 4.5 m
Smart Functions	Person counting, object/person localisation

2.2 Audio System

The main component in audio-based object localisation and tracking is the microphone array. We use a total of 64 microphones (56 microphones used in three microphone arrays for acoustic source localisation and the other eight microphones for recording of speech and music data). The scheme of technical connection between all audio components is illustrated in [3]. Each microphone array consists of microphones of solely the same model. There are two arrays consisting of 16 microphones each and one array consisting of 24 microphones. Figure 2 shows the three microphone arrays. The aim of using different types of microphones is to examine the influence of different microphone properties on the performance and accuracy of audio classification and localisation tasks within the scope of targeted use cases of the laboratory.

The technical specifications of used microphones are given in Table 2. The Nowsonic Calibration microphone as well as the Justin JM-714 are both omnidirectional microphones. The directional characteristic of the MXL 840 is cardioid. This makes the MXL 840 well suited for experiments where it might be beneficial to reduce or to focus on specific sounds coming from certain directions. The frequency ranges of the Nowsonic Calibration microphone and the MXL 840 correspond to the capabilities of human hearing. The Justin JM-714 by contrast has a limited top end. The physical properties (dimensions, diameter) make the Justin JM-714 suitable for possible integration into embedded systems (in our case in the Intenta S2000 sensor), whereas the other two microphones need an external mounting.

The arrays are all constructed from aluminium profiles which are arranged into rectangular grids. The individual microphones are mounted into these profiles using special plastic brackets. The array geometries can be modified. This is done by vertical as well as horizontal adjustment of the aluminium profiles and microphones. This flexibility allows for fast, simple, reliable and reproducible examination of the effects which different array geometries have on localisation and classification algorithms and their performance. Due to the physical dimensions of the aluminium profiles, the microphone diameters and the corresponding brackets, certain ranges of distances between microphones can be set up within each array. The horizontal and vertical range is 3.3 – 13.2 cm and 6 – 17.1 cm, respectively. In addition to adjusting the array geometry, the arrays as a whole can as well be flexibly positioned within the active tracking area. Each array can be flexibly mounted anywhere to the outer frame. This enables for audio recording from different positions and directions and thus enables for another set of experiments to investigate the respective impact on localisation and classification performance. Besides the fixed array-mounted microphones, there are a number of additional mobile microphones available for various related applications such as location/field recording of sounds or music. This is especially beneficial for compiling sound sets for research in the field of audio classification. The available microphones are: 2 × Audio Technica AT4040, 2 × Rode NT5-MP, 2 × BLX288/PG5 wireless handheld microphones, and 2 × BLX188/CVL wireless lavalier microphones.

Table 2 – Technical specifications of microphones and corresponding array configurations

	Nowsonic Calibration (Microphone array 1)	**MXL 840** (Microphone array 2)	**Justin JM-714** (Microphone array 3)
Type	electret condenser	pressure gradient electret condenser	electret condenser
Directionality	omni	cardioid	omni
Frequency range (Hz)	20 – 20000	30 – 20000	50 – 16000
Dimensions (mm)	21 × 200	22 × 134	9 × 28
Diameter (mm)	6	22	< 5
Number of microphones	16	16	24

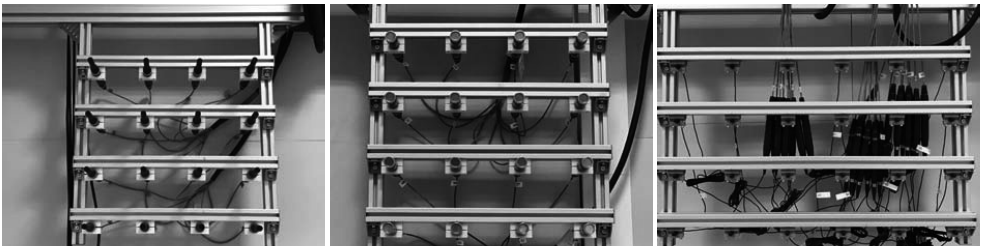

Figure 2 – The three microphone arrays (left to right: Nowsonic Calibration, MXL 840, and Justin JM-714)

Audio signals played back by loudspeakers guarantee for precise reproducibility of the experiments. Within the laboratory, there are 16 loudspeakers in total. The specifications of the 16 loudspeakers which are usable to mimick audio sources/objects are summarised in Table 3. The Genelec as well as Tannoy brand loudspeakers are well-recognised and industry standard loudspeakers. In order to investigate the influence of sound source properties on localisation and classification tasks, different models of loudspeakers had been chosen. For the investigation of the effect on amplitude-based localisation methods, different loudspeakers with different power and sound pressure levels are available. The influence of the sound source's frequency range can be examined with the help of the different available loudspeakers as well. Figure 3 (right) shows an example of configuration and positions of microphone arrays and loudspeakers within the laboratory.

Table 3 – Technical specifications of loudspeakers

	Genelec 8010 AP	**Genelec 8020 CPM**	**Genelec 8030 BPM**	**TANNOY Reveal 402**	**TANNOY Reveal 502**
Power (Watt)	25	20	40	50	75
Sound pressure level (dB)	96	105	108	101	108
Frequency range (Hz)	74 – 20000	66 – 20000	58 – 20000	56 – 48000	49 – 43000
Number of loudspeakers	2	8	2	2	2

As it has been described in [3], the backbone of the audio system is built by a recording/playback computer, audio interfaces and an audio master clock generator. The master clock generator we use is a ROSENDAHL NanoClocks GL. This is connected to all audio interfaces as well to the audio computer in a star topology. The audio interfaces we apply are 2 × Focusrite Octopre MK II Dynamic for connecting microphones 1 - 16 as well as all 16 loudpeakers. All other microphones are connected through 6 × Focusrite Octopre MK II audio interfaces. The signals of all 8 Focusrite interfaces are gathered in one RME ADI-648 interface, which is connected to the computer with an audio RME HDSPe PCIe MadiCard.

All audio data is recorded within the software Steinberg Cubase Pro 8.5 EDU. The employed sampling frequency is 48 kHz at a resolution of 24 bit per audio sample. The signal of one microphone is recorded into one mono audio channel. Thus, one recording contains 64 channels (for 64 microphones). Playback of audio data is also handled via Cubase which offers powerful routing and sequencing capabilities for controlling which loudspeaker plays back which signal at any given time. A further software which is employed and available to the users of the laboratory is the self-developed audio annotation tool [9].

In addition to the core elements, the audio systems comprises a number of extra peripheral devices. These include a PreSonus headphone amplifier, two Beyerdynamic DT-770 Pro headphones, a Roland A-88 MIDI controller keyboard and a Yamaha DGX-660 digital piano. The MIDI controller keyboard can be used as an input device to control Cubase's virtual instrument and sound generation capabilities. Thus, from triggering parasitic noises through imitating classic instrument sounds, the MIDI controller keyboard offers a well-known, human-accustomed, standard interface. For even increased haptic, playability and more direct access to sound generation without a computer, the Yamaha DGX-660 digital piano is available to classically trained musicians.

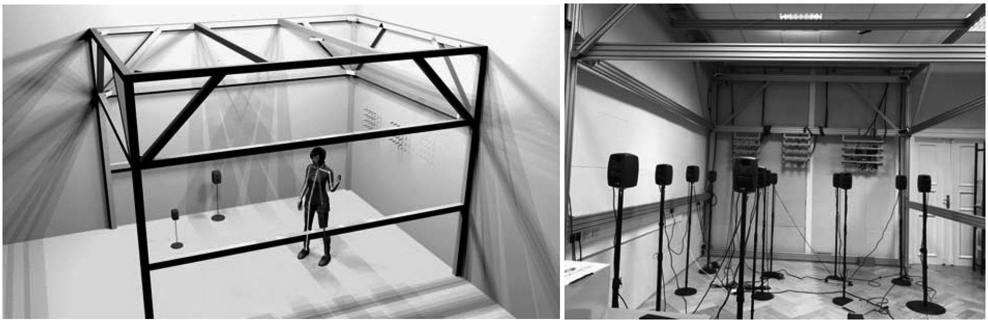

Figure 3 – Laboratory: (left) 3D model with two speakers and one person, (right) experimental setting with multiple speakers and three microphone arrays

3 Computational Analysis Tools and Internet of Things

The lab user gets access to a set of web-based annotation tool-kits allowing collective labelling with standard functionality for audio, video and image labelling, like bounding boxes, polygons, binary classifications, temporal segmentation, categories/keywords, free text captioning and many more. The annotation tools have been applied in previous TRECVid evaluation campaigns and are partially described in [10, 11]. Even virtual reality (VR) based annotation as recently evaluated is available [12]. Annotation can be further supported by pre-processing of data allowing to do annotations in correction mode. Three GPU workstations each one including an NVidia P6000 graphics card enable fast processing of recorded video footage e.g. with body keypoint based human pose estimation algorithms, illustrated at the person in Figure 3. Such model-driven annotation strategies support and accelerate the scenery analysis and evaluation of existing algorithms [13]. Additionally, profound access to IoT sensor probes for temperature, humidity, light intensity and more are provided through cooperation [14].

4 Conclusion and Future Perspectives

We introduced our new audio-video laboratory and its current main focus, which is the creation of systematic audio and video data aiming for development of new or evaluation of existing acoustic source localisation and tracking methods. Acoustic event classifications in indoor scenarios were recently demonstrated on data sets including several hundreds of actions using machine and especially deep learning algorithms [4, 5]. In the future, we like to extend the amount and diversity of such audio data, enrich them with annotated information and document scenarios with high precision by applying simultaneously multi-perspective video recordings and representing most, if not all, objects and persons in 3D or VR environments as function of space and time.

5 Acknowledgements

This work is funded by the program of Entrepreneurial Regions InnoProfile-Transfer in the project group *localizeIT* (funding code 03IP608).

References

[1] *Laboratory of Fraunhofer-Institut für Digitale Medientechnologie IDMT*. Available on `https://www.idmt.fraunhofer.de/de/hsa/services/equipment.html`, 2018. Last accessed at 19. January 2018.

[2] Richter, C., J. Lindemann, R. Römer, and M. Wolff: *Das Labor für kognitive Systeme an der BTU Cottbus-Senftenberg*. In *Elektronische Sprachsignalverarbeitung 2015 (ESSV 2015)*, no. 78 in Studientexte zur Sprachkommunikation, pp. 240–247. TUDpress, Eichstätt-Ingolstadt, 2015.

[3] Hussein, H., R. Manthey, A. Hasan, M. Heinzig, M. Ritter, D. Kowerko, and M. Eibl: *Design of a Laboratory for Audio and Video Based Object Localization and Tracking*. In *Proc. of International Summer School on Computer Science, Computer Engineering and Education Technologies (ISCSET)*. Laubusch, Saxony, Germany, 2017.

[4] Hussein, H., M. Ritter, R. Manthey, J. Schlosshauer, E. Fabian, and M. Heinzig: *Acoustic Event Classification for Ambient Assisted Living and Health Environments*. In *Elektronische Sprachsignalverarbeitung 2016 (ESSV 2016)*, no. Band 81 in Studientexte zur Sprachkommunikation, pp. 271–278. TUDpress, Leipzig, 2016. URL `http://www1.hft-leipzig.de/ice/essv2016/index_de.html`.

[5] Kahl, S., H. Hussein, E. Fabian, J. Schlosshauer, E. Thangaraju, D. Kowerko, and M. Eibl: *Acoustic Event Classification Using Convolutional Neural Networks*. In *INFORMATIK 2017*, pp. 2177–2188. Gesellschaft für Informatik, Bonn, Chemnitz, 2017. doi:10.18420/in2017_217. URL `https://dl.gi.de/handle/20.500.12116/3989`. DOI: 10.18420/in2017_217.

[6] Zietlow, T., H. Hussein, and D. Kowerko: *Acoustic Source Localization in Home Environments - The Effect of Microphone Array Geometry*. In *Elektronische*

Sprachsignalverarbeitung 2017 (ESSV 2017), no. 86 in Studientexte zur Sprachkommunikation, pp. 219–226. TUDpress, Saarbrücken, 2017. URL `http://essv2017.coli.uni-saarland.de/index.html`.

[7] KAHL, S., T. WILHELM-STEIN, H. HUSSEIN, H. KLINCK, D. KOWERKO, M. RITTER, and M. EIBL: *Large-Scale Bird Sound Classification using Convolutional Neural Networks*. In *CEUR Workshop Proceedings (Working Notes of CLEF 2017 - Conference and Labs of the Evaluation)*, vol. 1866. 2017. URL `ceur-ws.org/Vol-1866/paper_143.pdf`.

[8] *Blender (Software)*. Available on `https://www.blender.org/`, 2018. Last accessed at 21. January 2018.

[9] ROSCHKE, C.: *Entwurf und Implementierung eines webbasierten Managementsystems zur Entwicklung und Optimierung von Audio- und Videoanalysealgorithmen*. Master's thesis, Chemnitz University of Technology, Chemnitz, Germany, 2016.

[10] RITTER, M., M. HEINZIG, R. HERMS, S. KAHL, D. RICHTER, R. MANTHEY, and M. EIBL: *Technische Universitat Chemnitz at TRECVID Instance Search 2015. TRECVid Workshop Proceedings*, 2015. URL `http://www.researchgate.net/profile/Marc_Ritter2/publication/272831068_Technische_Universitt_Chemnitz_at_TRECVID_Instance_Search_2014/links/54f115890cf2f9e34efd4778.pdf`.

[11] KAHL, S., C. ROSCHKE, M. RICKERT, D. RICHTER, A. ZYWITZ, H. HUSSEIN, R. MANTHEY, M. HEINZIG, D. KOWERKO, M. EIBL, and M. RITTER: *Technische Universitat Chemnitz at TRECVID Instance Search 2016. TRECVid Workshop Proceedings*, 2016, pp. 1–8, 2016. URL `https://www.researchgate.net/publication/312211838_Technische_Universitat_Chemnitz_at_TRECVID_Instance_Search_2016`.

[12] KAHL, S., D. RICHTER, C. ROSCHKE, M. RICKERT, M. HEINZIG, D. KOWERKO, M. EIBL, and M. RITTER: *Technische Universität Chemnitz and Hochschule Mittweida at TRECVID Instance Search 2017. TRECVid Workshop Proceedings*, 2017, pp. 1–7, 2017.

[13] KOWERKO, D., D. RICHTER, M. HEINZIG, S. KAHL, S. HELMERT, and G. BRUNNETT: *Evaluation of CNN-based algorithms for human pose analysis of persons in red carpet scenarios*. In *INFORMATIK 2017*, pp. 2201–2209. Gesellschaft für Informatik, Bonn, Chemnitz, 2017. doi:10.18420/in2017_219. URL `https://dl.gi.de/handle/20.500.12116/3991`. DOI: 10.18420/in2017_219.

[14] KURZE, A., A. BERGER, and S. TOTZAUER: *A "Kinder" Surprise: Big Brother Is Watching You(r Humidity Values*. 2017. URL `https://www.researchgate.net/publication/318055199_A_Kinder_Surprise_Big_Brother_Is_Watching_Your_Humidity_Values`.