59th ILMENAU SCIENTIFIC COLLOQUIUM
Technische Universität Ilmenau, 11 – 15 September 2017
URN: urn:nbn:de:gbv:ilm1-2017iwk-089:0

# HUMAN CAPTURE SYSTEM FOR VIRTUAL REALITY APPLICATIONS

*H. Drumm*

Technische Universität Ilmenau, Kompetenzzentrum Virtual Reality, Germany

## ABSTRACT

The growing competition between the manufacturing companies results in constant optimization of their processes. This applies not only to the production process, but also to the process of product development. Digitization plays a very important role in this development.

Through the process of digitization, things can be quickly designed, changed, and visualized. One of the great innovations that helps visualize the process of three dimensions is virtual reality (VR). With virtual reality, you are able to test and improve simulated processes interactively with user help [PD0915].

In this paper, an experimental hardware and software system is to be presented, called "***Human Capture System***". This system allows for integration of various data sources like sensors or simple data streams.

***Index Terms –*** Virtual Reality, sensors, simulation, interaction

## 1. INTRODUCTION

In recent years, the use of virtual reality in our lab at the university has shown that there are some existing deficiencies with the professional VR-system. One of the shortcomings is the lack of possibility to record the users during the experiments. Our VR laboratory was designed only to visualize and make audible content. Therefore, it was not possible to check the results of the experiments in retrospect.

Another missing feature was the use of new technologies for interaction, e.g. voice recognition, body tracking, or gestures. Today's VR-applications require the use of very different user data for designing interactive events. Based on our experiences, we not only have tried to develop a system which attempts to monitor and record the user data, but also to use the data for the interactivity of the scene. In this paper, an experimental hardware and software system is to be presented, called "***Human Capture System***". This system allows for integration of various data sources like sensors or simple data streams.

## 2. TECHNICAL POSSIBILITIES

The first step is to describe the technical possibilities for the system. In order to record the users during the experiment, a device is necessary, which provides image and audio data. If you want to capture users from different perspectives, then a camera array or clip-on microphone is necessary. For practical use, it is important to record the timecode, e.g. to prevent delays between audio and video.

The head is the part of the body that provides the most information for evaluation methods. Image analyses is used to recognize person-specific data, such as gender, age, or direct identification per se. The extraction of this kind of data from images has already been done through the use of neural networks [LZXW14, HYYK15, TYRW14]. The extraction of person-specific data to be collected or recorded for VR applications can be useful for the creation and access of user profiles. People can create or even access saved profiles through facial recognition, and can set specific features for the content automatically, e.g. gender and age.

The detection of emotions from image data is already used in many non-professional cameras such as the so-called "smile sensor". The recognition of emotions can be used to control or to trigger events in VR-applications and can also be evaluated for media psychological examinations.

For facial recognition, a large image section must represent the face in order to support a precise evaluation. For this purpose, a high-resolution camera should be used [Pana17].

The eye movements can also be tracked using special glasses [KE2008]. This generates the center of attention for the user in relation to the scene content; however, it must not be included in the interactive control of a scene.

Face tracking can be used to transfer facial expressions to avatars, e.g. in the gaming or movie industry. The classical face-tracking technique uses markers to recognize facial movements. Face-tracking without such markers is the result of the combination of image processing and neural networks.

The speech recognition has already found its way into the daily life of many people, e.g. accessing a mobile phone in a car, or by controlling a TV. On the other hand, it is not commonly used in the VR sector. There are great potentials for controlling VR scenes with regard to navigation, but it also aids in the evaluation of speech in the calculation of reactions in virtual avatars [PD15].

An extension of speech recognition is used to determine a person's intonation. Intonation being the flow or emphasis of a person's speech. From the data, speech recognition can also be used for psychological analyses.

The measurement of EEG waves is a tried and tested method for detecting disturbances of brain function. In science, the EEG waves of humans are evaluated with the aid of pattern recognition in order to control devices, e.g. prosthetics. This technique can also be used to control VR applications [FLPS10].

Now we consider the entire body, which delivers data for biological feedback. Skin resistance, blood pressure, and heart rate are a few examples that are important for determining the degree of excitement or anxiety. This can serve both to evaluate the psychological condition as well as to control the scene content of the VR scene.

There are several technical solutions for the realization of body tracking [AC99, MHK06]. Body tracking is most commonly used with markers, but other technical solutions include the use of magnetic fields, ultrasound, or image recognition by means of neural networks [Kinect13]. Body tracking is mainly used for the transmission of human movement to animate avatars, but it can also be used for the evaluation of body gestures [VP17]. For instance, through specific arm positions in conjunction with the "pointing gesture" you can see a person's intensions of pointing at an object in a certain direction.

Gesture recognition without markers is based on image recognition by means of neural networks. BMW was one of the first companies to take gesture recognition towards industrial applications, e.g. under the subject of Industry 4.0 at the BMW Group [FBMW16].

For finger tracking, the two technical solutions are marker-free and marker-bearing. Marker-free tracking enables more freedom to move, but according to our experience this method is not stable for use in the professional field. In virtual prototyping, finger tracking is the basis for an improved immersion in the handling of 3D models. The goal of improved immersion in VR is the natural movement of a human through virtual scenes.

The 3D scanning of objects is nothing new [RHL02]. What is new is that the depth of data can be recorded quickly and in a resolution that plays a role in industrial applications. Autonomous robots capture their surroundings by means of depth cameras or lasers and orient themselves in the excess space [SNH03]. Similar techniques include recording a person's body for collision detection in the virtual environment. 3D depth detection devices create point clouds, which can be used for collision detection. A more advanced processing step is the calculation of the geometric structure. This process describes the surface properties (e.g. skin color) so that a scanned body can be completely reconstructed. 3D-scanning of users is useful for teleconferencing systems or to simply see one's own body in VR. This is important while using head-mounted displays, since the contact to the real environment is completely blown out [BF17, BF15].

## 3. SYSTEM OVERVIEW

The following figure summarizes a selection of technologies already present, which have been partly converted into our hardware and software applications. Our strategic objectives of the development have been to create a hardware- and software system which is capable of integrating several technologies. The measured and processed data is used for controlling the content in the VR-applications as well as for observing the user during the experiment.

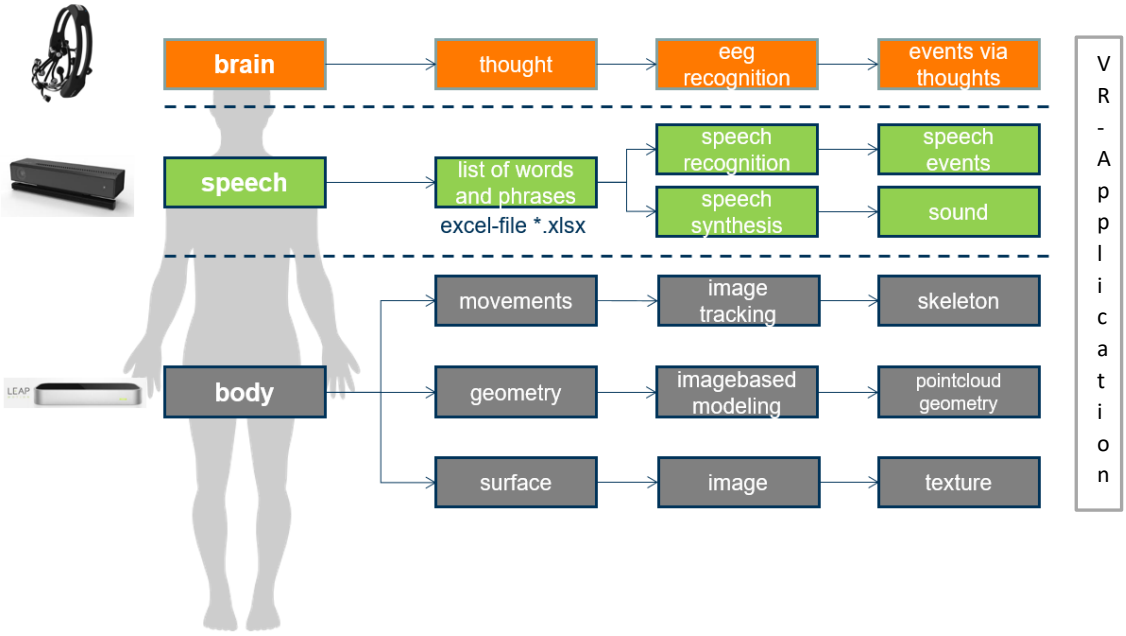


*Figure 1: Systematic overview about the data mining, processing and resulting data*

Figure 1 shows the flow from the input sensors at the left side to the resulting processed data at the right. Different sensor types capture the human output.

Furthermore, it is shown that creation technologies are used to provide resulting data, e.g. speech recognition, body tracking, etc. On the right side of figure 1, the final results are calculated from the measured data, e.g. speech events or a human skeleton. The data is extracted as visual data, audio data, tracking data, EEG waves, coordinates and more.

## 4. HARDWARE STRUCTURE

Figure 2 shows the hardware structure of the developed system. It consists of a client-server system. Each client is able to integrate one or multiple sensors, e.g. Microsoft Kinect V2 Sensor, one Leap Motion Sensor, or the combination of both. The limitation of the number of sensors per client results from the device requirements, the amount of measured data, and the caused calculation effort.

Currently the human capture system operates with Microsoft Kinect V2 Sensors [Kinect13] and Leap Motion Sensors [Sk13]. The sensor for thought control has not yet been implemented, but it would be possible to extend the system. The Kinect V2 Sensor has been used because it satisfies various requirements at once and also supports, video-, audio-, depth-data, gestures, body and face-tracking, and speech recognition. The Leap Motion Sensor is state of the art and uses marker-free finger tracking.

The server works as a control center to manage the system, e.g. to add, remove, and configure clients. The clients are connected to the server via the infiniband network (40 Gbit) created by Mellanox [MX].

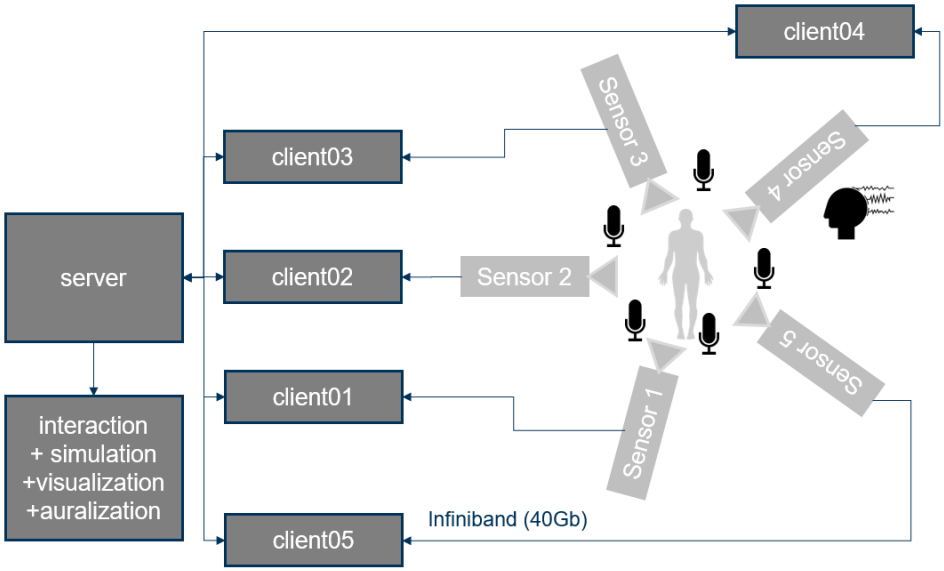


*Figure 2: Systematic overview about the hardware structure*

## 5. SOFTWARE MODULES

The resulting software contains different modules. One of the modules contains the loading, creating, saving, and managing of the system configuration profiles (see figure 3). Up to eight clients with connected hardware are definable in a system profile. The profile consists of a master and client IP address and the associated sensors or devices to be used. While the system is running, the resulting data is then sent through the network to the server.

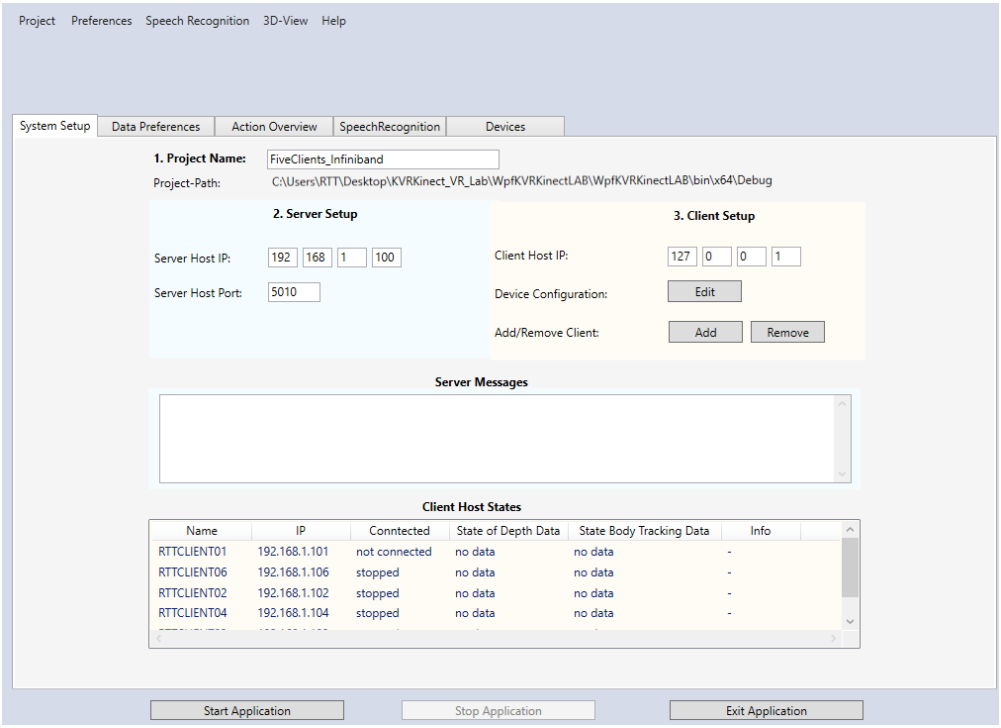


*Figure 3: Module for loading, creating, saving, and managing of the system configuration profiles*

Another module includes the calibration process. If you want to use tracking data, you have to combine the individual sensors in a common coordinate system. In this case, the user selects which reference coordinate system to use, for example, an external tracking system (e.g., ART). It is also possible for the user to select an internal reference coordinate system of the sensor. The software responsible for a

very simple calibrating the external tracking system was realized in the master thesis of Jan Kurtz [Kz16]. Final calibration corrections can also be made with graphical tools in the client software.

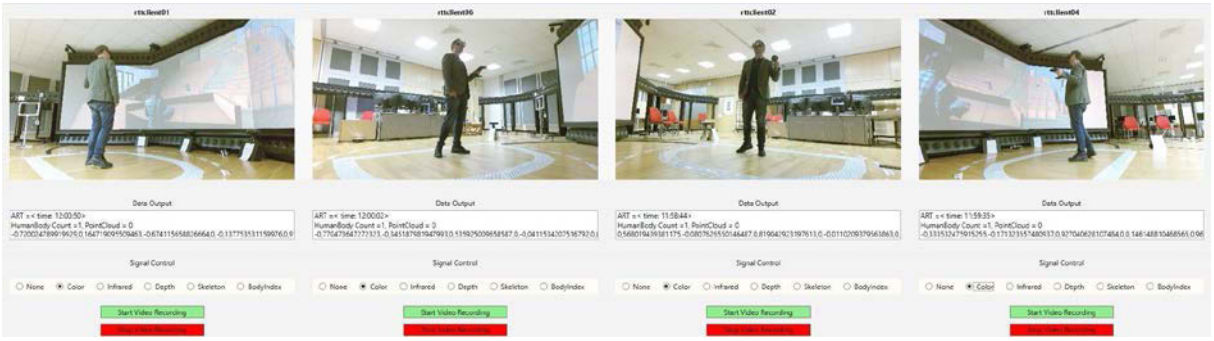

*Figure 4: Subject monitoring*

The next topic of the software is subject monitoring (Figure 4). It is possible to select and record various with "pan", "tilt" and "zoom" is conceivable. However, camera control is not possible with the Kinect V2 sensor due to its lack of a remote control. For this purpose, is better to use the [Pana17].

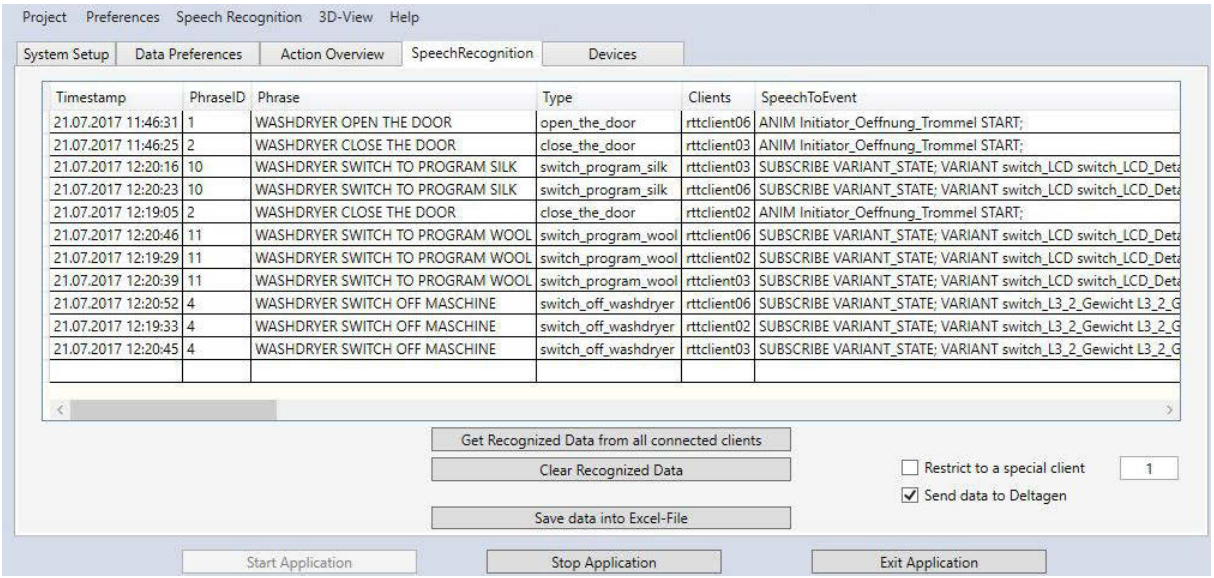


*Figure 5: Module for speech recognition*

Another module of the software deals with the evaluation of the measured tracking and voice data. One part of the module contains the speech recognition (Figure 5). The user of the software can create their own excel list of words or sentences for speech recognition in order to load them into the module. The same list can then be used for commands that control the external software which triggers certain speech events. A speech event monitor regulates the speech events with their properties, e.g. recognized words, sentences, timecode, and client names produced by the individual sensors. This list of events can be saved for later evaluation. Through the use of the Kinect V2 sensor, the first experiments have shown that speech recognition only works well if no flow text is used. The speech recognition also requires a microphone array in order to evaluate speech from several directions.

*Figure 6: Markerless skeleton tracking and pointclouds from different views*

Currently, up to five Kinect V2 sensors are used to perform body tracking. With the calibration of the sensors, it is possible to superimpose these skeletons (Figure 7). However, the superimposition is not yet precise since the optical distortions of the sensors are not taken into account. Experiments were conducted to use tracking with gesture recognition for the professional field in the CAVE. The system is still not completely accurate and requires additional algorithms to stabilize the tracking data.

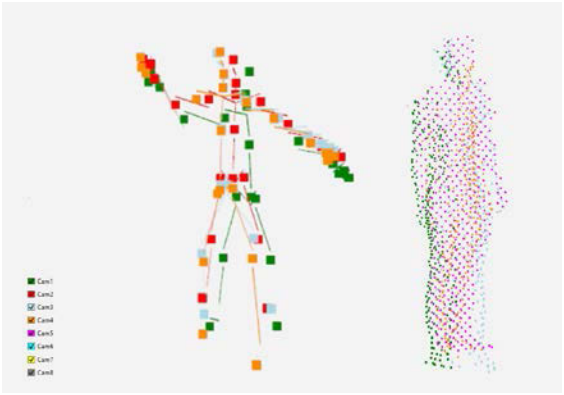

*Figure 7: Superimposition of 5 skeletons and 5 pointclouds*

## 6. CONCLUSIONS

At present, software is still in the prototyping phase, and the functions have to be improved and adapted to our constantly changing needs.

## REFERENCES

[PD0915] Pöschl, S., & Döring, N. (2015, September). *User Experience in Virtual Reality Application Development – Design and Evaluation of a Fear of Public Speaking Scenario.* Paperpräsentation auf der 9. Konferenz der Fachgruppe Medienpsychologie Division der DGPs, Tübingen.

[LZXW14] Wei Li, Rui Zhao, Tong Xiao, Xiaogang Wang, Chinese University of Hong Kong. *DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification*, CVPR 2014.

[HYYK15] Guosheng Hu, Yongxin Yang, Dong Yi, Josef Kittler, William Christmas, Stan Z. Li, Timothy Hospedales, *When Face Recognition Meets with Deep Learning: an Evaluation of Convolutional Neural Networks for Face Recognition.* ICCV Workshop 2015 on IEEE Conference.

[TYRW14] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: *Closing the gap to human-level performance in face verification. In Computer Vision and Pattern Recognition (CVPR)*, 2014 IEEE Conference on, pages 1701–1708. IEEE, 2014.

[Pana17] Panasonic 2017, http://business.panasonic.de/professional-kamera/remote-kameras/integrierte-pt-kameras/aw-ue70

[KE2008] Kai Essig, Vision-Based Image Retrieval (VBIR): A New Eye-Tracking Based Approach to Efficient and Intuitive Image Retrieval. VDM Verlag Dr. Müller, ISBN-10: 3836492415, 2008

[PD15] Poeschl, S., & Doering, N. (2015, June). *Measuring Co-Presence and Social Presence in Virtual Environments - Psychometric Construction of a German Scale for a Fear of Public Speaking Scenario.* Paper presented at the 20. Annual CyperPsychology und CyperTherapy (CyPsy20), La Jolla, California.

[FLPS10] Doron Friedman, Robert Leeb, Gert Pfurtscheller, Mel Slater, *Human–Computer Interface Issues in Controlling Virtual Reality With Brain–Computer Interface*, journal: Human–Computer Interaction,Vol. 25, Nr. 1, pp. 67-94, 2010, http://www.tandfonline.com/doi/abs/10.1080/07370020903586688

[AC99] Jake K. Aggarwal, Qin Cai, *Human motion analysis: a review Computer Vision and Image Understanding (CVIU)*, 73 (3) (1999), pp. 428-440

[MHK06] Thomas B. Moeslund, Adrian Hilton, Volker Krüger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding, Volume 104, Issues 2–3, 2006, pp. 90-126

[VP17] Radu-Daniel Vatavu, Smart-Pockets: Body-deictic gestures for fast access to personal data during ambient interactions, International Journal of Human-Computer Studies, Volume 103, 2017, Pages 1-21, ISSN 1071-5819, http://dx.doi.org/10.1016/j.ijhcs.2017.01.005.

[Kinect13] T. Hanna, *Microsoft KINECT: Programmierung des Sensorsystems Taschenbuch*, Sept 2013, ISBN-13: 978-3864900303.

[Sk13] M. Spiegelmock, *Leap Motion Development Essentials*, Oct 2013, ISBN-10: 1849697728.

[FBMW16] Speech Klaus Fröhlich at the BMW Group Press Conference CES 2016, https://www.press.bmwgroup.com

[RHL02] S. Rusinkiewicz, O. Hall-Holt, M. Levoy, *Real-Time 3D Model Acquisition.* SIGGRAPH '02 Proceedings of the 29th annual conference on Computer graphics and interactive techniques Pages 438-446

[SNH03] H. Surmann, A. Nüchter, J. Hertzberg, *An autonomous mobile robot with a 3D laser range finder for 3D exploration and digitalization of indoor environments.* In: Robotics and Autonomous Systems Volume 45, Issues 3–4, 31 December 2003, Pages 181-198.

[BF17] S. Beck, B. Froehlich, *Sweeping-Based Volumetric Calibration and Registration of Multiple RGBD-Sensors for 3D Capturing Systems.* In Proceedings of IEEE VR 2017, Los Angeles, USA, March 2017, pp. 167-176.

[BF15] S. Beck, B. Froehlich, *Volumetric Calibration and Registration of Multiple RGBD-Sensors into a Joint Coordinate System.* In Proceedings of IEEE Symposium on 3D User Interfaces (3DUI), Arles, France, pp. 89-96, March 2015. DOI=10.1109/3DUI.2015.7131731

[MX] http://www.mellanox.com/

[Kz16] J. Kurtz, TU-Ilmenau, Department Medientechnolgie, Master Thesis: *Nutzerzentrierte Entwicklung einer Kalibriersoftware für Kameras im Virtual Reality Lab*. 2016

## CONTACTS

Dr.-Ing. H. Drumm helge.drumm@tu-ilmenau.de