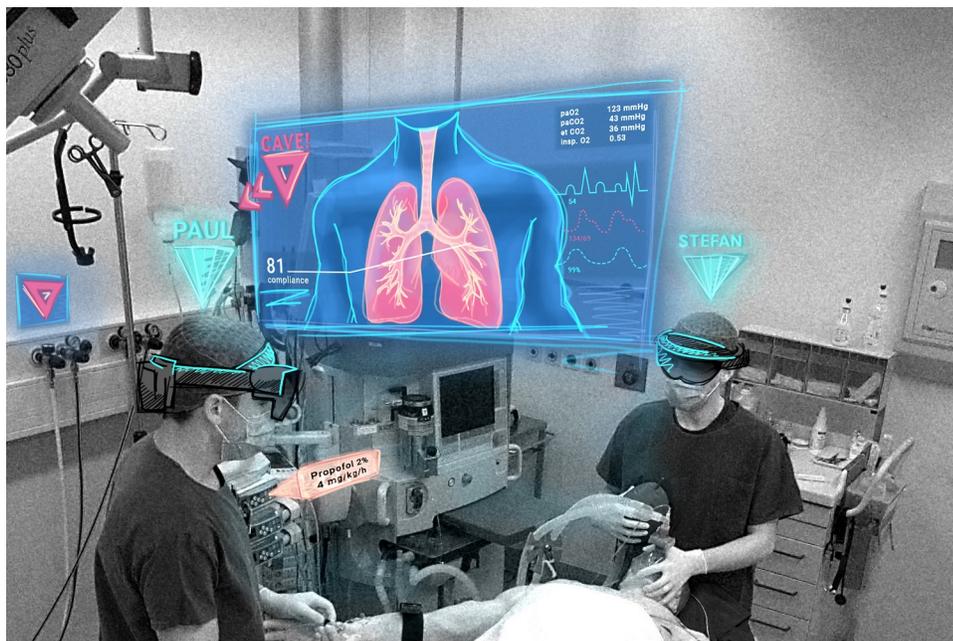


---

# Exposé

<Student Names>  
<Student IDs>

---



Project <Title>

Supervisor: Prof. Dr. Florian Niebling  
Summer Term 2023

<Date>

# 1 Motivation and Goals

Interactions with keyboard or mouse do not fulfill human capabilities and seem unnatural sometimes. New and natural ways of Human-Computer-Interaction (HCI) is lacking. Therefore, mid-air (also in-air) interactions seem to be promising input modalities since they are easy, quick to use, intuitive and natural. What gestural systems should be used always depends on a particular context. In a medical context, anesthetists have to deal not only with bacteria transmission, but also with several medical devices monitoring the patients' health condition.

Thus, a project that aims to create a pervasive interaction with the help of augmented reality in cooperation with the University Hospital Würzburg. Touch screens will be prepared to replace the screen of an anesthesia machine to enable recognition of remote input on the interface of the machine.

The following HCI project aims to create intuitive gesture detection for these AR applications in a medical simulation room. It is crucial to decide on an interaction method to communicate and interact with the AR application to increase user experience. My part in this process is to figure out an intuitive set of hand gestures and how to learn and operate them reliably. For this project, AR technology like the HoloLens 2 will be used.

Part of the standard functionalities brought by the Microsoft HoloLens are its featured gestures. These consist of "Touching", "Raycasting", "Focussing", and the "Air Tap". Under the use of these specific gestures, multiple contextual interfaces, like Pop-Ups, have been designed for HoloLens experiences. However, the limited amount of gestures face the problem of limited interaction possibilities and lower the variety of controls. The idea is now, to expand the repertoire of available gestures and therefore expand the interaction and control possibilities to create a pervasive interaction with the system.

The main goal of this project is to learn and detect new gestures using a machine learning framework, Unity Engine and the HoloLens 2 by Microsoft. This goal can be separated into several sub goals and come with various decision making processes:

Decision on the gesture set, finding and implementing the right network for the selected gestures to learn with and implementing a Unity-HoloLens interface for hand tracking. In this expose, suggestions for possible implementations are made. The actual decisions will be made in the following procedure of the project after the acceptance and agreement on the project contents.

## 2 Related Work

### 2.1 Hand Gesture Recognition Studies

Few to no research has been done in the field of HoloLens gesture expansion. Therefore, first literature concerning hand gesture recognition in general was selected.

The first study used Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) for skeleton based human activity and hand gesture recognition (*Núñez et al. 2017*). Their learning process was based on 3D posture and vision based action recognition problems based on a combination of a CNN and LSTM. Experimental testing showed that their developed training method obtains better results than a single-stage training strategy. Furthermore, they suggested data augmentation and emphasized that they received the best results using small data sets where the proposed data augmentation strategy has greater impact.

The second paper to point out is called "Skeleton-based Dynamic hand gesture recognition" (*Smedt et al. 2016*). It proposes an approach for 3D hand gesture recognition. They also focused on the geometric shape of the hand to extract effective descriptors from hand skeletons returned by a depth camera. They evaluated an approach containing 14 gestures. The results showed significant higher performance over a depth-based approach.

In addition to that, another study addresses a 3D gestural interface (*Katherina A. Jurewicz et al. 2018*). The researchers compared gesture-function mappings between experts and novices and discovered both similarities and differences between the groups. The system they used was 3D, vision based and required gestural input in an operating room in the context of anesthesia.

### 2.2 Gestural Interaction using Egocentric Sensing

Another research paper considering egocentric sensing of mid-air gestural input was found (*Katherina A. Jurewicz et al. 2018*). In this particular study, they used a head-mounted depth camera in a smart space "office" environment. To publish hand tracking information, they developed a C++ application using different kind of libraries. They examined multiple HMD-based gesture techniques and detected a high preference of mid-air gestures compared to head gestures for input methods.

### 2.3 Gesture Elicitation Studies

Besides on deciding on how to implement the skeleton-based hand gesture recognition, deciding what gestures to use is also a crucial point.

Multiple Gesture Elicitation Studies or Gesture Elicitation Review Studies specifically covering Mid-Air Interactions have pointed out that there are no established, universal gesture vocabularies for mid-air interactions themselves and explain that "each input method is best at something and worse at something else" (*Villarreal-Narvaez et*

al. 2020, Vogiatzidakis and Koutsabasis 2018, Katherina Ann Jurewicz 2020). They primarily emphasized the fact, that the identification of appropriate gestures depends on the context of use and are very important design decisions for a system. and have been present since the late 70s. The research paper states out that gestures should be selected by specific criteria, like discoverability, memorability, performance, reliability and comfort. Moreover, they also pointed out that it is very important to figure out the gesture "appropriateness", like focusing on "easy to perform", inutivite or learnability.

## 3 Concepts

### 3.1 Unity-HoloLens Interface

An interface between the Unity Engine and the HoloLens to provide hand tracking and dynamic data has to be created. This can be most likely done with the Mixed Reality Toolkit (MRTK). Instructions on how to use the MRTK with the HoloLens are directly given by Microsoft on their documentation website. Additionally, with this Toolkit it is possible to create a hand tracking profile to generate joint prefabs (see Figure 1). Therefore, relevant joints for gesture detection can be selected and used for further operations. During this process I have to be very cautious with the complexity of the scripts, since joint objects are transformed on every frame and can have significant performance cost. This leads to the assumption to sort out non-relevant joints.

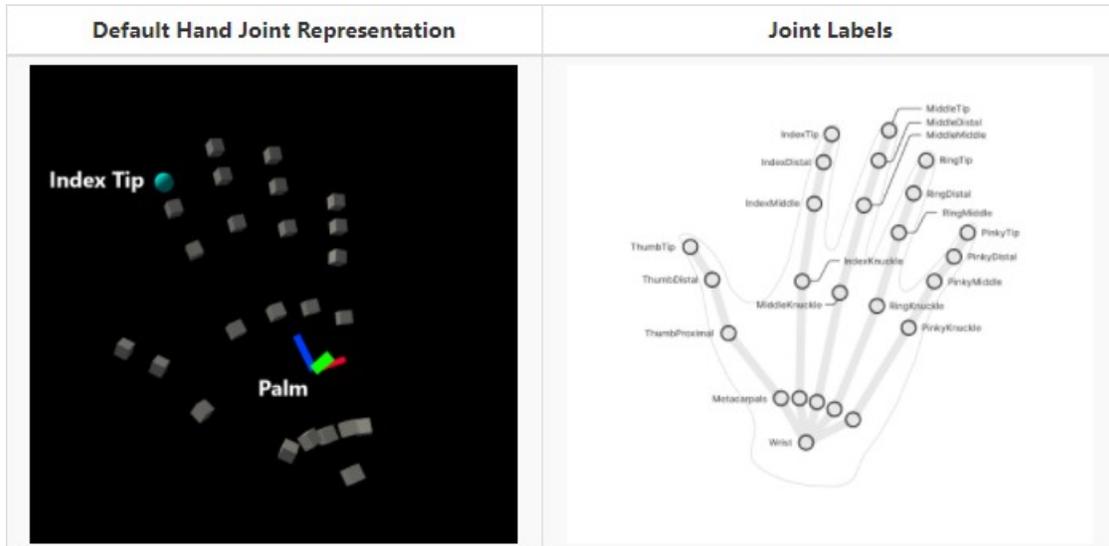


Figure 1: Hand Tracking with MRTK (*microsoft.github 2020*)

### 3.2 Potential Neural Networks

After reading some related work, more than one machine learning frameworks come into question. In this paragraph, these frameworks will be introduced. I will not be able to make a final decision on the used framework until I've spent some time getting to know the associated hardware. However, the final decision will be made on quality aspects like reliability or complexity.

The first framework that comes to mind and that is often referenced in the literature is the Convolutional Neural Network (CNN). This network consists of three different types of layers: The Convolutional Layers, the Pooling Layers and the Fully-Connected Layers. Often, more than one level of Convolutional-Pooling Layers exist. Application areas of CNNs are often Image Processing, but also Object Detection and Face Recognition. The primary task of the network is to classify content or identify objects in the scene (nose, eyes, ears, etc) and cluster them.

Secondly, the Recurrent Neural Network (RNN) can be found in the literature. This kind of network is distinguished by its cycles. Therefore, previous outputs can be reused as inputs in the new cycle to make it possible to remember previous states. RNNs are often used for pattern detection, but also find use in the section of image processing, video tagging or face detection.

Another possible network is the Long Short Term Memory Network (LSTM Network). The LSTM is a type of RNN which is capable of learning long-term connections. The key to LSTM is the cell states which can be described as a conveyor belt running horizontally through the diagram. It runs down a certain chain with only linear interaction while being controlled by certain gates. It is very easy for information just to flow along. A visualized comparison of RNN and LSTM can be found in Figure 2 and Figure 3.

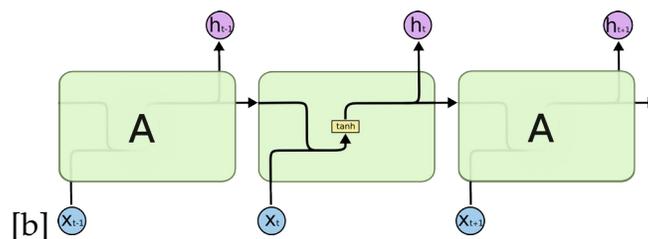


Figure 2: Repeating module in a standard RNN contains a single layer (Carter 2020)

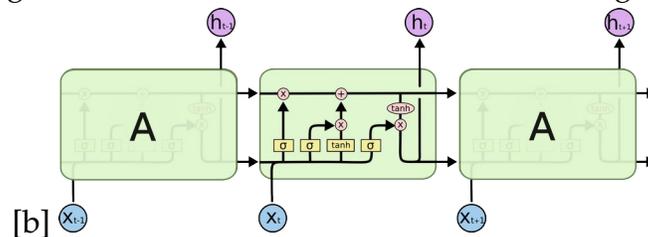


Figure 3: Repeating module in an LSTM contains four interacting layers (Carter 2020)

Besides all the previously mentioned deep learning networks, a Support Vector Machine (SVM), a linear or non-linear model for classification and regression problems, is also conceivable. Compared to other machine learning methods SVM is very powerful at recognizing patterns and complex data sets. Application areas of SVMs are among other things recognizing fraudulent credit cards, identifying speakers, as well as detecting faces.

## 4 Methodology

Multiple decisions have to be made during the development of this project. Some of the main decisions will be mentioned in this paragraph.

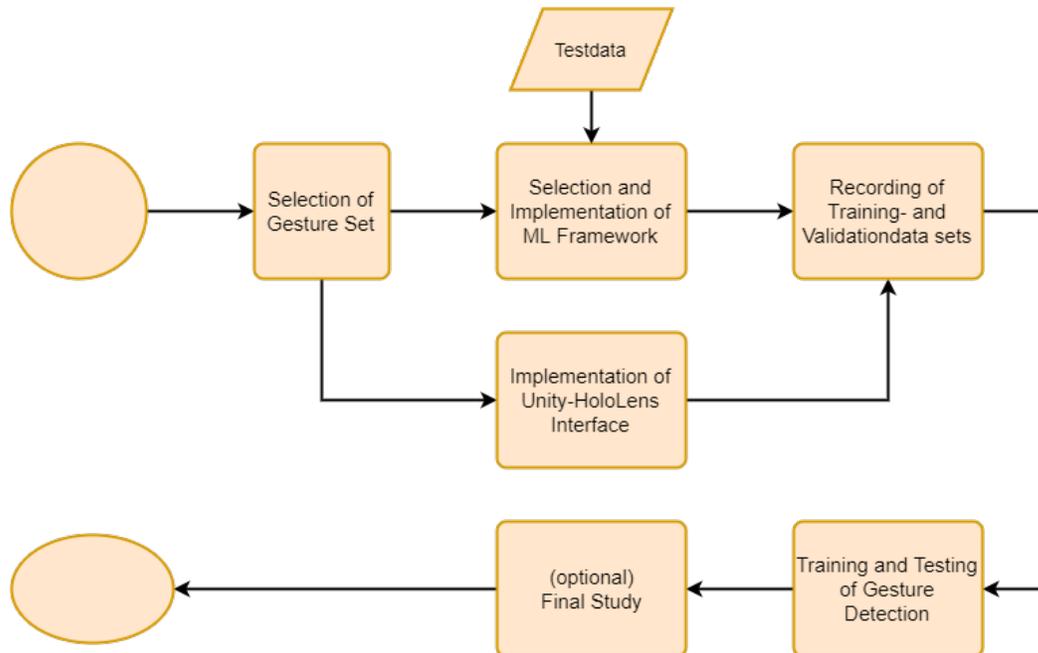
The main focus of this project is to learn and perform mid-air gestures from an egoperspective using the HoloLens 2. First and foremost, a decision on the hand gesture set has to be made. Whether the literature will be sufficient to select the gestures or if an elicitation study has to be considered will be decided after finishing the complete literature research.

Next, preparations for the deep learning process have to be made. This means creating the Unity-HoloLens interface and recording training or test data on the one hand, and on the other hand it means implementing and adapting the Machine Learning Framework. Possible interface implementations are already discussed in the section Concept, as well as possible Machine Learning Frameworks.

After creating a functional Unity-HoloLens interface and a Network that is capable of learning, a set of training and validation data has to be recorded and transferred to the Machine Learning Network. This data will be used for the actual training process of the network. Afterwards, the learning results will be used and tested for gesture detection, hopefully using the HoloLens 2 already.

If there is still time after this procedure, a study can be executed to test and verify the quality of the gestures and the final results of the project.

## 5 Sequence of Operation Diagram



## 6 Schedule

1. Literature Analysis (1 week)  
Analyse different mid-air gestures for interaction
2. Gesture Set generation if no third-party data can be used, preparation of data from literature otherwise (2 weeks)
3. Analysis of deep learning process (2 weeks)  
Recording and preparation of training set data and process
4. HoloLens Interface creation (1 week)
5. Initial training and test for gesture recognition on device (1 weeks)
6. User study preparation (2 weeks)  
Questionnaires, statistical analysis
7. Evaluation and examination of empirical data from user study (2 weeks)
8. Writing of report (2 weeks)

## References

- Carter, R. (2020). *Exploring gesture control for ar applications*. <https://www.uctoday.com/unified-communications/exploring-gesture-control-for-ar-applications/> (accessed: 18.10.2020)
- Jurewicz, K. A. [Katherina A.], Neyens, D. M., Catchpole, K., & Reeves, S. T. (2018). Developing a 3d gestural interface for anesthesia-related human-computer interaction tasks using both experts and novices.
- Jurewicz, K. A. [Katherina Ann]. (2020). Using a bayesian framework to develop 3d gestural input systems based on expertise and exposure in anesthesia.
- microsoft.github. (2020). *Hand tracking*. <https://microsoft.github.io/MixedRealityToolkit-Unity/Documentation/Input/HandTracking.html> (accessed: 18.10.2020)
- Núñez, J. C., Cabido, R., Pantrigo, J. J., Montemayor, A. S., & Vélez, J. F. (2017). Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition.
- Smedt, Q. D., Wannous, H., & Vandeborre, J.-P. (2016). Skeleton-based dynamic hand gesture recognition.
- Villarreal-Narvaez, S., Vanderdonckt, J., Vatavu, R.-D., & Wobbrock, J. O. (2020). A systematic review of gesture elicitation studies: What can we learn from 216 studies?
- Vogiatzidakis, P., & Koutsabasis, P. (2018). Gesture elicitation studies for mid-air interaction: A review.