A deep learning approach for gesture-to-text in Augmented Reality

Junxiao Shen[a,b]

[a] *1Department of Engineering, University of Cambridge, UK*
[b] *js2283@cam.ac.uk*

Keyword: Human-Computer Interaction, Deep Learning, Machine Learning, Gesture Text Input, Augmented Reality

1. Research Goal

Due to the hardware inability to accurately sense the trace of the figure, there is a big displacement between the trace of the figure tips we see in the user interface in HoloLens and the actual data input of the trace. Therefore, a recognition model to map the trace to word or sentences should be invariant of the starting position of the keyboard. Therefore, Deep learning can be used to train this recognition model. Innovated from speech recognition, since speech is a 1D temporal data and the trace is a temporal and spatial data, therefore, the network models and state-of-art algorithms for speech recognition can be used to innovate us to design and train the neural networks for the trace recognition.

2. Research Plan

Figure 1 shows the overall plan. it is a component diagram and it is used to visualise and help us to have a better idea of the working process.

2.1. Data Synthesis

We will thus use the data we collected from the users' experiment to synthesis the data together with the data for the gaze collected from the experiment since we also believe that the trace of the gaze can be relevant to the final output as well. And the synthesised data will be from used to train the neural network and the trained model will thus replace the recognition component in Figure 1. The reason to synthesis the data is that unlike speech recognition, the recorded labelled data can be accessed quite easily nowadays since people can pay others to read and label whereas in our project. The dataset is limited due to /the lack of funding to hire a lot of people to generate the traces and due to the short time frame considering it is an only one-year project.

2.2. Data Training

We will do another users experiment of the established system and test the results and compare them with the state-of-art text input system. The user experience will still be carried out in the OptiTrack system.

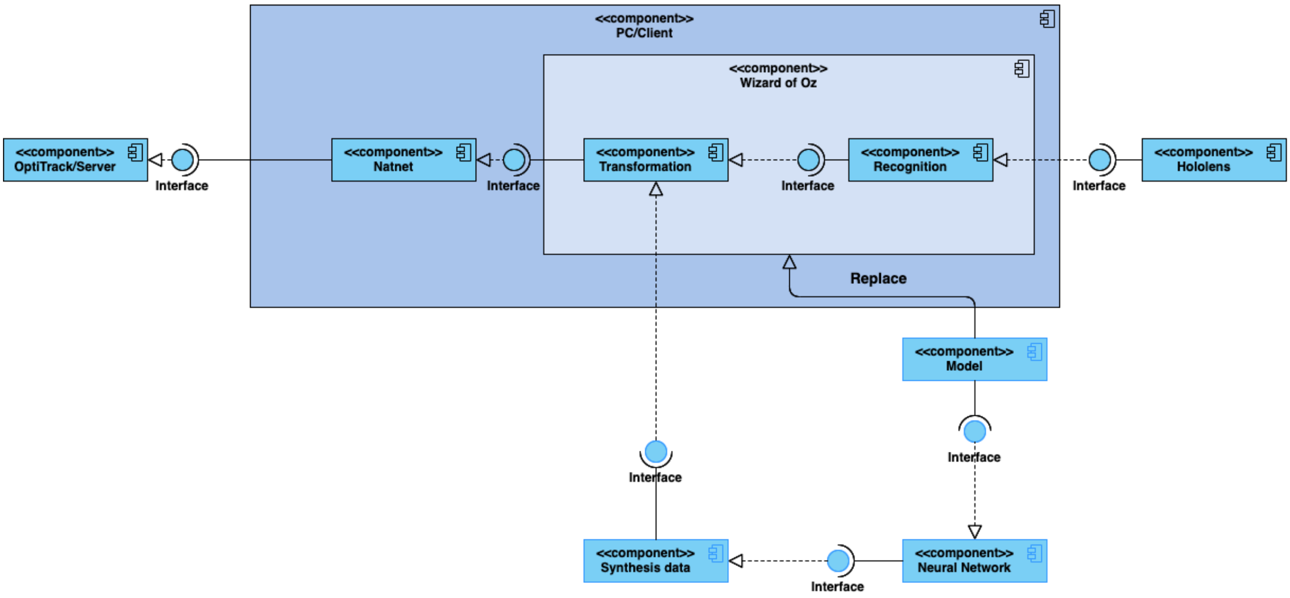

Figure 1: Component diagram for the development

Different neural networks can be used. I would start from Recurrent Neural Networks (RNN) first since the trace of fingertips is a 3d spatial and temporal signal and this can be in analogy to speech recognition. Furthermore, we could also try to combine CNN and RNN to see whether it will lead to better results. Transfer learning can also be used since the fine-tuning train ml model with synthetic data in first iterations and the with real data in further iterations. We should also be careful that we shouldn't use synthetic data in a test set since it can have a corresponding pair of real data which may lead to a dangerous result.

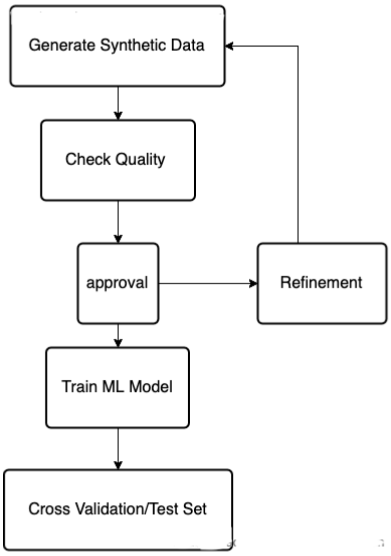

Figure 2: work flow for second and third stage

3. Cost

The majority of the cost here will be for the computational power used to generate the data and to train the deep neural networks. We are going to use Google Cloud to run the neural network and doing prototyping and large scale production code on this platform consistently can be very expensive. An alternative is to buy a strong GPU straight away and then the front cost is higher but the marginal costs is much lower. However. regardless of which way to use, 1 thousands pounds will be the starting point for this machine learning intensive project. Therefore the cost in our case is simple and straight away.