REACTION PAPER

In times where robots such as “LocoMan” are being developed in order to aid human in different condition including old age, human gesture learning will definitely help incorporate virtual humans, virtual assistants, robots, metahumans into the human society. As highlighted in the talk gesture forms a very important part of human communication, it helps express more thoughts and emotions, enhance and regulate communication. A major advantage being it helps improve user engagement and rapport. This evaluation has not been a major point of concern, where the user engagement with the robot/virtual human was judged based on the rapport build between the user and the robot/virtual human. Considering the practical application of virtual humans being used in the therapy, without gestures it would not be as effective as an actual therapist to builds a rapport with the patient to understand the patients’ mental condition and provide solutions in an effective manner. This difference was well demonstrated by the speaker with an animation of a virtual human with just audio and the one with gesture. The difference in impact is huge.

Another significant feature of the research presented was the baseline for the judgement of the correct gesture. This needs to consider various facts about gesturing which was rightly highlighted by the speaker which is there is no 1:1 mapping between words uttered and the gesture presented. It could be controlled by various factors which could be as absurd as that of the energy of the person speaking. At the end of the day, the same words might not draw extreme gestures when compared that at the beginning of a fresh day. Some of the words might invoke different emotions leading to different gestures in different people which is again a study of psychology and sociology. Defining a baseline for such wide horizon of input would be difficult task, but essential since it would help evaluate the findings objectively. In one the animations presented by the speaker the outcome was much different from that of the real video and yet considered a presentable output. This would be of concern if the judgement is clearly objective. The pipeline structure where the words uttered are converted to gestured need to pass through a filter of emotions. If a sentiment of the statement does not invoke higher degree of emotions, then the gestures are limited and vice versa. The tokenization concept which was discussed by the speaker might be tuned to include emotions too.

The work presented also included the gestures and did not account for the facial expressions which go hand in hand. Gestures without facial expression almost take away half the impact. Though the speaker mentions that fact that, including facial expressions might be a distraction, which can be addressed by making an objective evaluation of the gestures and not subjective which is currently being made. Thought the quantitative evaluation was discussed, there is need to limit the subjective evaluation to major factors. In any real world application, weights associated with facial expression would be higher when compared to bodily gestures, and in combination the effect would be much more practical.

From the presentational perspective, the flow of the talk was well organized where the speaker first gave an idea as to how data is widely available and can be used to model or learn different objectives which included gesture learning, 3D scene reconstruction. The speaker present the existing work in the field by dividing them into different categories such as motion capture, extracting gesture from wild videos. This is followed by mentioning the drawbacks and how it is rectified in the speakers work. The quantitative analysis very well highlighted the gain over other methods, and even slight improvement was highlighted by the speaker when the comparison or

error was between low pass filter and the model. Subjective evaluation was well documented through animations which was a useful tool to capture the attention of the audience. The speaker also mentioned where the idea was inspired from, example GenAI was used for text-to-gesture synthesis which was quite intuitive. The speaker also demonstrated a practical application where due to the innate nature of gesture where there is 1:many mapping, current models average out the input data to provide the output, which was rectified where the model does token classification instead of regression. Instead of the mathematical analysis, a subjective analysis was the right way to convey the issue to the audience. Another interesting case was the demo of the output from the low pass filter which removed even the necessary gestures with the intent of making the output less jittery. These demonstrations helped convey the issues to the audience in a better manner than just with equations. Even in the case of modeling the 3D environment, the speaker differentiates between the cases where it is are already optimized such as making the gaussian to be as closer to the image as possible with every step, and the ones which need optimization such as gaussian does not cover all the detail in the picture which is corrected using Atomized proliferation.