Abstract:

Virtual agents will be defined in this paper as a software program which uses scripted rules and, increasingly, machine learning and artificial intelligence in order to provide automated service or guidance to humans. Virtual agents may be more or less feature rich, including capabilities such as speech, animation, etc. The animation of virtual agents is an important field of research. Animation allows for more realistic mimicry of real-world behaviour. It allows users to interact with virtual agents more intuitively and naturally, as it more accurately portrays human-human interaction.

The simplest approach to generating animation in virtual agents is simple hand animation. A skilled worker crafts each animation from their own study of human behaviour. This method yields high-quality results but is limited by the availability of skilled workers and is very time intensive. It may also lead to a lack of variation in the agent’s movement capabilities, as only a set number of animations will be made.

A machine learning approach seeks to generate similarly high-quality expressive animation without the costly and time-intensive downsides of traditional animation.

Deep learning-based motion synthesis methods can probabilistically generate any number of different convincing output animations based on natural language audio inputs, which yields greater variation in gestures than traditional hand-animation.

A transformer-based architecture trained on a data set of aligned natural speech and motion capture data may be employed to interactively generate expressive gestures for virtual agents by supplying natural language text inputs.

In this project, we explore different methods of synthesising gestures by treating it as a sequence-to-sequence problem. We implement a Transformer-based motion synthesis engine which, supplied with motion capture data of a single actor’s performance, as well as a transcription of their speech, can generate new gestures when presented with novel text input data.