Towards a Data Generation Framework for Affective Shared Perception and Social Cue Learning Using Virtual Avatars*

Matthias Kerzel[1] and Stefan Wermter[1]

[1]Knowledge Technology, Department of Informatics, University of Hamburg, Germany
kerzel / wermter @informatik.uni-hamburg.de
http://www.knowledge-technology.info

1 Extended Abstract

Research on machine learning models for affective shared perception, social cue, and crossmodal conflict learning generates a high demand for large data sets of accurately annotated and unbiased training samples While many existing data sets rely on freely available "in-the-wild" video material or paid actors, using fully controlled virtual avatars has a series of advantages: 1) Once scripted, virtual avatars and environments can be automatically varied and randomized to generate any desired number of training samples. 2) Generated video material can be automatically annotated with the exact time point of avatar behavior, e.g., exact information about the gaze target, the position of hands and body pose, obviating the tedious hand-annotation process. 3) The generated behavior is fully controllable, allowing a detailed analysis of the contribution of different behaviors to machine learning and participant study results. 4) Full control over biases, e.g., actor appearance and positioning in a scene can be controlled and balanced, unwanted behavior can be excluded.

To this end, we suggest a fully scriptable framework for shared perception and social cue learning with highly-realistic virtual avatars, realized in Blender[1] and Python, for easy integration with established machine learning toolkits. The framework is centered around a table-scenario with up to 4 simulated persons (avatars), with optional objects on the table, that can serve as a focus for shared perception. Furthermore, robot models can be directly integrated and simulated within the framework for embodied and cognitive robots. The behaviors of the avatars include: different facial expressions as affective cues; realistic gaze shift as a combined eye and head motion, different types of gestures like pointing, social signaling, and beat gestures during talking; talking-animations; body pose for social cues and idle-animations for more realistic and life-like appearance; finally spatialized sound is included for sound source-based attention mechanisms. The

* The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169)

[1] Free and open source 3D creation suite Blender: https://www.blender.org/