Multimodal Human Machine Interaction in Virtual Reality

Jean Bernard, Gerard Chollet, Annie Gentes, Patrick Horain, Catherine Pelachaud, Dianle Zhou & Leila Zouari

Virtual worlds are developing rapidly over the internet. They are visited by avatars and staffed with Embodied Conversational Agents (ECAs). An avatar is a representation of a physical person. Each person controls one or several avatars and usually receives feedback from the virtual world on an audio-visual display. Ideally, all senses should be used to feel fully embedded in a virtual world. Sound, vision and sometimes touch are the available modalities.

Myblog 3D research focuses on diversifying the characteristics of our "portraits" and augmenting our interactions. The purpose is to go beyond mere information and to enlarge the palette of emotions, the subtlety of expressions, and the diversity of relations. It also takes into consideration the environment of the avatars and ECAs as relevant in the process of mutual understanding. We aim to offer new paradigm to control one's own avatar living in a 3D environment. The virtual environment is a 3D blog; that is a virtual environment populated with virtual agents and objects. One can invite friends for a chat, exchange of documents such as music, photos, and so on. In most systems involving avatars, users ought to type in what s/he has to say, select a scripted animation to move his/her avatar. It is thus cumbersome to animate one's own avatar. In the project MyBlog-3D, the avatar's animation is driven by the user's behaviour in an unobtrusive manner. A small webcam detects the arm movements as well as the facial expressions a user displays. These detected motions are reproduced in real-time onto the avatar. On the other hand, the ECA animation is driven by a multimodal dialog system.

This project addresses three axes of research: face and gesture processing, speech recognition, embodied conversational agent and we question their potential in creating significant interactions within 3D environments. In the first axe, we study the relation between the face texture and the parameter gradient matrix. We propose a statistical approach to fit the AM to unknown users by dynamically estimating the gradient and update matrices from the face texture. We have implemented this algorithm for real time face tracking and demonstrate its robustness when tracking multiple or unknown users. Regarding the second axe of research, speech recognition, we are elaborating a system to segment the continuous speech stream into speech / non-speech. We are also developing a keyword spotting algorithm trained on the ESTER database. Detecting keyword allows us to get 'important' information for the ECA to have a 'sense' of what is being said. Doing a semantic analysis is out of reach for a real-time application and when the discourse domain is simply too vast. Keyword spotting allows us to get some information related to the conversation and allows the ECA to answer using a template-based dialog system. The output of the dialog is communicative intentions. These intentions are then converted into multimodal nonverbal behaviours.

Our aim is not only to enhance technological aspects of our algorithm but also to evaluate the added-value of our technologies within the 3D Blog application. We have defined a scenario of self training using the combination of these technologies. Users design their avatar and meet with an ECA, a Human Resources Director, for a job interview. They can observe and learn through practicing with the ECA how to behave in such circumstances. We rely on users habits to test themselves and play with aspects of their personality in relatively secure environments and on the development of job interviews in virtual worlds such as Second Life.