Mimicking from Perception and Interpretation

Catherine Pelachaud[1], Elizabetta Bevacqua[1], George Caridakis[2], Kostas Karpouzis[2], Maurizio Mancini[1] , Christopher Peters[1], Amaryllis Raouzaiou[2]
[1]University of Paris 8
[2] National Technical University of Athens
c.pelachaud@iut.univ-paris8.fr

The ability for an agent to provide feedback to a user is an important means for signalling to the world that they are animated, engaged and interested. Feedback influences the plausibility of an agent's behaviour with respect to a human viewer and enhances the communicative experience. During conversation addressees show their interest, their understanding, agreeing and attitudes through feedback signals. They also indicate their level of engagement. It is often said that speaker and addressee dance with each other when being engaged in a conversation. This dancing together is partly due to the mimicking of the speaker's behavior by the addressee. In this paper we are interested in addressing this issue: mimicking as a signal of engagement.
We have developed a scenario whereby an agent senses, interprets and copies a range of facial and gesture expression from a person in the real-world. Input is obtained via a video camera and processed initially using computer vision techniques. It is then processed further in a framework for agent perception, planning and behaviour generation in order to perceive, interpret and copy a number of gestures and facial expressions corresponding to those made by the human. By perceive, we mean that the copied behaviour may not be an exact duplicate of the behaviour made by the human and sensed by the agent, but may rather be based on some level of interpretation of the behaviour (Martin et al., 2005). Thus, the copied behaviour may be altered and need not share all of the characteristics of the original made by the human.

General framework

The present work takes place in the context of our general framework (Figure 1) that is adaptable to a wide range of scenarios. The framework consists of a number of interconnected modules. At the input stage, data may be obtained from either the real world, through visual sensors, or from a virtual environment through a synthetic vision sensor.

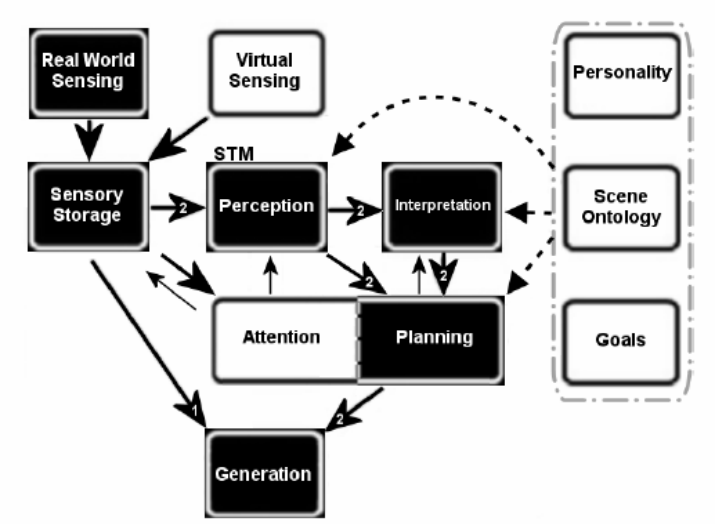

Fig. 1. The general framework that embeds the current scenario. Large arrows indicate the direction of information flow, small arrows denote control signals, while arrows with dashed lines denote information availability from modules associated with long term memory. Modules with a white background are not applicable to the scenario described in this paper.

Visual input is processed by computer vision (Rapantzikos and Avrithis, 2005) or synthetic vision techniques (Peters, 2005), as appropriate, and stored in a short-term sensory storage. Gesture expressivity parameters, facial expressions are extracted from the input data (Bevacqua et al., 2006). This acts as a temporary buffer and contains a large amount of raw data for short periods of time. Elaboration of this data involves symbolic and semantic processing, high-level representation and long-term planning processes. Moreover, it implies an interpretation of the viewed expression (e.g. value of Facial Animation Parameters (FAPs) extracted → anger), which may be modulated by the agent (e.g. display an angrier expression) and generated in a way that is unique to the agent (anger → another set of FAPs or of FAPs values). The generation module (Hartmann et al., 2005) synthesises the final desired agent behaviours.

Application scenario

Currently our system is able to extract data from the real world, process it and generate the animation of a virtual agent. Either synthesized gesture or facial expression are modulated by the gesture expressivity parameters extracted from the actor's performance. The input coming from the real world is a predefined action performed by an actor. The action consists of a gesture accompanied by a facial expression. Both, the description of the gesture and of the facial expression are explicitly requested to the actor and previously described to him in natural language (for example the actor is asked "to wave his right hand in front of the camera while showing a happy face"). The *Perception* module analyses the resulting video extracting the expressivity parameters of the gesture and the displacement of facial parts that is used to derive the FAPs values corresponding to the expression performed. The FAPs values and the Expressivity parameters are sent to the *Interpretation* module. If the facial expression corresponds to one of the prototypical facial expression of emotions, this module is able to derive its symbolic name (emotion label) from the FAPs values received in input; if not the FAPs values are used. Instead, the symbolic name of the gesture is sent manually because the Interpretation module is not able to extract the gesture shape from the data yet. Finally, how the gesture and the facial expression will be displayed by the virtual agent is decided by the *Planning* module that could compute a modulation either of the expressivity parameters or of the emotion. Then the animation is calculated through the Face and the Gesture Engine and displayed by the virtual agent.

References

Bevacqua, E., Raouzaiou, A., Peters, C., Caridakis, G., Karpouzis, K., Pelachaud, C. and Mancini, M. (2006), Multimodal sensing, interpretation and copying of movements by a virtual agent, Perception and Interactive Technologies, Kloster Irsee, June 2006

Hartmann, B., Mancini, M., and Pelachaud, C. (2005). Implementing expressive gesture synthesis for embodied conversational agents. In Gesture Workshop, Vannes.

Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., and Pelachaud, C. (2005). Levels of representation in the annotation of emotion for the specification of expressivity in ecas. In Int. Working Conference on Intelligent Virtual Agents, pp 405–417, Kos, Greece.

Peters, C. (2005). Direction of attention perception for conversation initiation in virtual environments. In Int.Working Conference on Intelligent Virtual Agents, pp 215–228, Kos, Greece.

Rapantzikos, K. and Avrithis, Y. (2005). An enhanced spatiotemporal visual attention model for sports video analysis. In International Workshop on content-based Multimedia indexing (CBMI), Riga, Latvia.