Real-time analysis and synthesis of emotional gesture expressivity

Maurizio Mancini[1] and Ginevra Castellano[2]

[1] University of Paris 8, Paris, France
m.mancini@iut.univ-paris8.fr
[2] University of Genova, Genova, Italy
ginevra.castellano@unige.it

Abstract. In this paper we focus on modeling a bi-directional communication between an embodied conversational agent and humans based on the non-verbal channel of the communication involving movement and gesture. We present a system that takes video data in input, extracts movement characteristics, and finally synthesizes the animation of a virtual agent. For each gesture performed by a human, the agent responds with a gesture that exhibits the same quality, i.e., the same characteristics of movement. We describe a mapping between the expressive cues we analyse in humans and the corresponding expressive parameters of the agent.

1 Introduction

In human-computer interaction, the ability for systems to understand users' behaviour and to respond to them with an appropriate feedback is an important requirement for generating an affective interaction. Systems must be able to create a bi-directional communication with users: analyzing their verbal and non-verbal behaviour to infer their emotional states and using this information to make decisions and/or plan an affective, empathic response. Virtual agent systems represent a powerful human-computer interface, as they can embody characteristics that a human may identify with and may therefore interact with the user in a more affective manner [18]. In this paper we focus on modeling a bi-directional communication between an embodied conversational agent and humans based on the non-verbal channel of the communication involving movement and gesture. Specifically, our research considers movement and gesture expressivity as a key element both in understanding and responding to users' behaviour. We present a system that takes video data in input, extracts movement characteristics, and finally synthesizes the animation of a virtual agent. Our system is made by the integration of two different software platforms: EyesWeb (www.eyesweb.org) [4] for video tracking and analysis of human movement, and the Greta embodied conversational agent (ECA) for behaviour generation [17]. In this scenario, we model a bi-directional communication based on real-time analysis of movement and gesture expressivity and generation of copying expressive behaviour: for each gesture performed by a human, the agent responds with a gesture that exhibits the same quality, i.e., the same characteristics of movement. We describe a mapping between the expressive cues we analyse in humans and the correspondent expressive parameters of the agent. Moreover, any information about the *shape* (configuration of hand/arm during the gesture execution) of the gesture performed by the user is completely ignored. Instead, only one particular gesture is reproduced by the agent to respond to the user's input.

In the next section we report the state of the art in the field and an overview of our system. Then we describe the extraction of expressive cues from a video input and the generation of expressive behaviour of the virtual agent.

2 State of the art

In the human-computer interaction field, a central role is played by automated video analysis techniques aiming to extract and describe physical characteristics of humans and use them to infer information related to the emotional state of individuals. Several studies focus on the relationships between emotion and movement qualities, and investigate expressive body movements ([20, 23, 8, 2]. Nevertheless, modeling emotional behaviour starting from automatic analysis of visual stimuli is a still poorly explored field. Camurri and colleagues ([5, 3, 7]) classified expressive gesture in human full-body movement (music and dance performances) and in motor responses of subjects exposed to music stimuli: they identified cues deemed important for emotion recognition and showed how these cues could be tracked by automated recognition techniques. Other studies show that expressive gesture analysis and classification can be obtained by means of automatic image processing [9, 1] and that the integration of multiple modalities (facial expressions and body movements) is successful for multimodal emotion recognition [11]. Several systems have been proposed in which virtual agents provide visual feedback/response by analysing some characteristics of the users' behaviour. In such systems the input data can be obtained from dedicated hardware (joysticks, hand gloves, etc), audio and video sources. SenToy [16] is a doll with sensors in the arms, legs and body. According to how the users manipulate the doll, they can influence the emotions of characters in a virtual game: depending on the expressed emotions, the synthetic characters perform different actions. Taylor et al. [21] developed a system in which the reaction of a virtual character is driven by the way the user plays a music instrument. Kopp et al. [14] designed a virtual agent able to imitate natural gestures performed by humans using motion-tracked data. Reidsma and colleagues [19] designed a virtual rap dancer that invites users to join him in a dancing activity. Users' dancing movements are tracked by a video camera and guide the virtual rap dancer in his own dance movements and gestures. In other previous works the Greta virtual agent was used to respond to external inputs. Mancini et al. [15] designed a system obtained by connecting emotion recognition in musical execution with Greta. In previous works [12] we have introduced the notion of expressivity for virtual agents. Starting from studies on the relation between emotional states and personality traits with the visual perception of the quality of human behavior [23, 10] the expressivity of the Greta agent is defined over a set of six dimensions, that modify the animation of the agent qualitatively. In this paper we describe the first application where the Greta agent responds to external motor behaviour analyzed in real-time.

3 System description

We present a system that allows us to analyse in real-time human movement and gesture expressivity and to generate expressive behaviour in an ECA. The system integrates two different platforms: EyesWeb [4] for video tracking and movement analysis and the Greta agent for behaviour generation [17]. Figure 1 shows an overview of the system architecture and modules:

- *Cues extraction*: we perform the automatic extraction of movement cues from a video source by using the EyesWeb platform. Further details on this part of the process will be given in Section 4.
- *Expressivity mapping*: in order to be able to send the extracted motion cues to the Greta agent, some mapping is needed. Motion cues are mapped into the agent's expressivity parameters and re-scaled. Further description of such process is given in Section 5.
- *Expressive engine*: this module is part of the Greta agent animation system. It receives expressivity parameters from the blackboard and computes the animation data needed to animate the virtual character. Starting from the results reported in [23], we have defined and implemented a set of expressivity parameters [13, 17] that affect the gestures (performed with hands/arms) quality of execution. For the purposes of the presented work, we turn our attention on four of these parameters: *Spatial Extent*, that changes the amplitude of movements (e.g., expanded versus contracted); *Temporal Extent*, that modifies the duration of movements (e.g., quick versus sustained actions); *Fluidity*, that determines the smoothness and continuity of movement (e.g., smooth, graceful versus sudden, jerky); and *Power*, that alters the dynamic properties of the movement (e.g., weak/relaxed versus strong/tense).
- *Visualization*: it is the Greta graphical engine that, given the animation data in input, creates a graphical representation of a virtual human that plays the animation back.

We designed a blackboard communication system in which data is exchanged between the two components through a blackboard structure, implemented with Psyclone [22]. All the system modules are connected to the same Psyclone blackboard via a TCP/IP socket as shown in Figure 1.

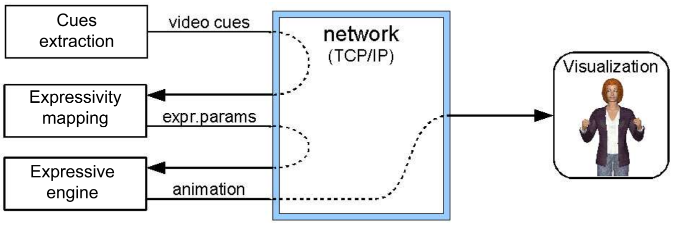

Fig. 1. Overview of the system architecture.

4 Automated extraction of expressive cues

Tracking of the full-body and the hands of the actors from the visual inputs was done in EyesWeb [4]. The EyesWeb Expressive Gesture Processing Library [6] was used for the automated extraction of the motion expressive cues. We analyzed both global indicators, from the full-body movement of the actors, e.g. the use of the space, and the dynamics of the barycenter of the hand used in the gesture. We extracted the following motion cues:

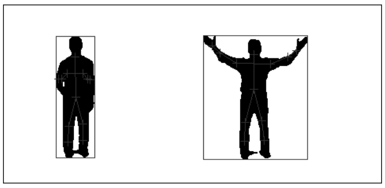

Fig. 2. High (left) and low (right) values of CI.

- *Contraction Index (CI)*: it is a measure, ranging from 0 to 1, of how a movement/gesture is contracted (i.e., performed near to the body) or expanded (i.e., performed with a wide use of the space surrounding the body). It can be calculated using a technique related to the bounding region (see Figure 2), i.e., the minimum rectangle surrounding the dancers body: the algorithm compares the area covered by this rectangle with the area currently covered by the silhouette. CI is a global indicator.

$$|v| = \sqrt{v_x^{\;2} + v_y^{\;2}}\text{, where } \begin{cases} v_x(t) = \dfrac{x(t+\Delta t) - x(t)}{\Delta t} \\ v_y(t) = \dfrac{y(t+\Delta t) - y(t)}{\Delta t} \end{cases}$$

Fig. 3. Velocity calculation.

- *Velocity*: we extracted the coordinates (x, y) of the barycenter of the hand used in the gesture and we calculated the module of the velocity, taking into account its horizontal and vertical components, as shown in Figure 3.
- *Acceleration*: we calculated the module of the acceleration of the hand used in the gesture, starting from module of velocity, as shown in Figure 4.
- *Directness Index (DI)*: it is a cue related to the geometric features of a movement trajectory. It is a measure of how much a trajectory is direct or flexible. In this context we applied the DI to the barycenter of the hand used in the gesture and we

$$|a| = \sqrt{a_x{}^2 + a_y{}^2} \text{ , where } \begin{cases} a_x(t) = \dfrac{v_x(t+\Delta t) - v_x(t)}{\Delta t} \\ a_y(t) = \dfrac{v_y(t+\Delta t) - v_y(t)}{\Delta t} \end{cases}$$

Fig. 4. Acceleration calculation.

considered it as an indicator of fluidity. In our implementation the DI is computed as the ratio between the length of the straight line connecting the first and last point of a given trajectory and the sum of the lengths of each segment constituting the given trajectory. Therefore, the more it is near to one, the more direct is the trajectory.

5 Expressivity copying

In this section we describe how we model the bi-directional communication between humans and agent based on movement and gesture. We defined the following correspondence between motion cues automatically extracted and the Greta expressivity parameters (Figure 5): the Contraction Index is mapped onto the Spatial Extent, since they provide a measure on the amplitude of movements; the Velocity onto the Temporal Extent, as they refer to the velocity of movements; the Acceleration onto the Power, as both are indicators of the acceleration of the movements; the Directness Index onto the Fluidity, as they refer to the degree of the smoothness of movements.

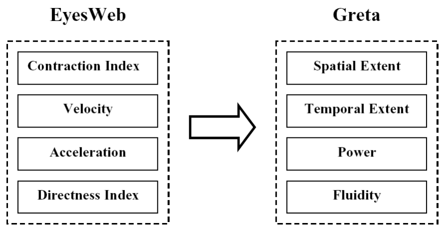

Fig. 5. Mapping between motion cues extracted with EyesWeb and the Greta expressivity parameters.

Since the motion cues and the expressivity parameters vary in different ranges, we performed a re-scaling (Figure 6). The indicators automatically extracted with EyesWeb (maximum of Velocity and Acceleration, minimum of the Contraction Index and average value of the Directness Index) were re-scaled depending on the overall minimum and maximum values of the correspondent motion cues calculated for each actor in all the emotional expressions.

The bi-directional communication between human and agent is based on real-time analysis of movement and gesture expressivity and generation of expressive copying

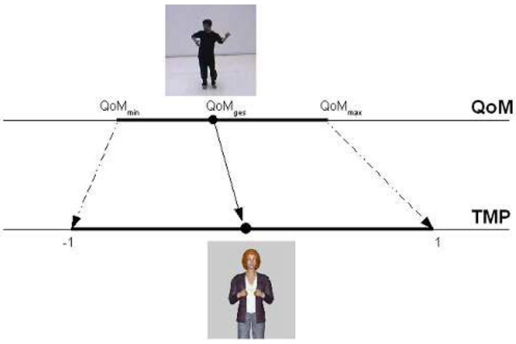

Fig. 6. Rescaling of indicators.

behaviour: for each gesture performed by the human, the agent responds with a gesture that exhibits the same quality, i.e., the same characteristics of movement. That is, any information about the *shape* (configuration of hand/arm during the gesture execution) of the gesture performed by the user is completely ignored. Instead, we have chosen a particular gesture (the one shown in Figure 7) which is the only one reproduced by the agent to respond to the user's input.

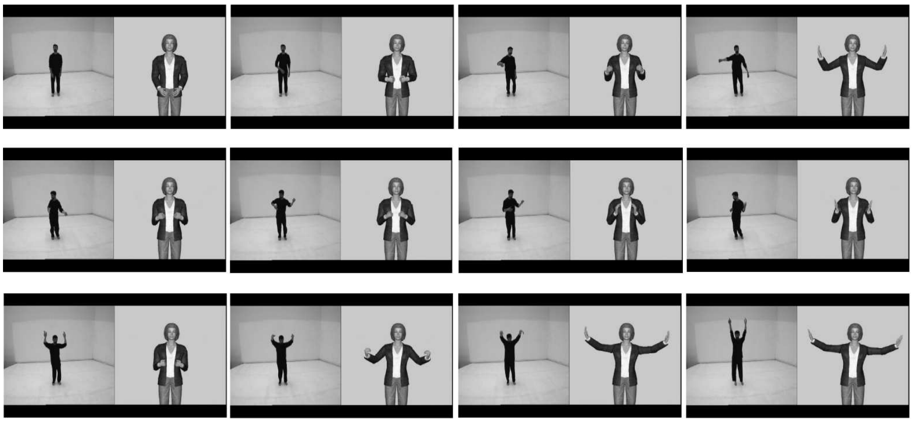

Fig. 7. Example of copying of expressive cues.

The indicators we chose to drive the synthesis process have proven successful: the expressive behaviour generation demonstrates that the synthesized gestures reproduce the movement expressivity of the real ones. Figure 7 shows some frames of the video that was provided as input to our system.

At the following link there is an example of motor behaviour generation:

`http://www.iut.univ-paris8.fr/~mancini/clips/acii07.avi`

6 Conclusions

In this paper we presented a system able to acquire input from a video camera, process information related to the expressivity of human movement and generate expressive copying behaviour. From the analysis perspective, we defined successful indicators to drive the synthesis process: the synthesized gestures reproduce the expressivity of the real ones. The real-time mapping between motion cues and expressivity parameters is direct, not involving emotion classification based on movement analysis.

Ongoing work focuses on providing our system with the real-time emotion mapping and automatic choice of a specific gesture depending on the emotion detected. We are planning perceptive tests with human participants that will allow us (1) to evaluate our synthesis approach based on real data analysis and (2) to improve both the analysis algorithms providing expressive cues and the synthesis approach for the generation of gesture expressivity in an embodied conversational agent.

7 Acknowledgements

The research work has been realized in the framework of the EU-IST Project HUMAINE (Human-Machine Interaction Network on Emotion), a Network of Excellence (NoE) in the EU 6th Framework Programme (2004-2007).

References

1. T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. Emotion analysis in man-machine interaction systems. In Hervé Bourlard Samy Bengio, editor, *Machine Learning for Multimodal Interaction*, volume 3361 of *Lecture Notes in Computer Science*, pages 318–328. Springer Verlag, 2005.
2. R. Thomas Boone and Joseph G. Cunningham. Childrens decoding of emotion in expressive body movement: the development of cue attunement. *Developmental Psychology*, 34:10071016, 1998.
3. A. Camurri, G. Castellano, M. Ricchetti, and G. Volpe. Subject interfaces: measuring bodily activation during an emotional experience of music. In J.F. Kamp S. Gibet, N. Courty, editor, *Gesture in Human-Computer Interaction and Simulation*, volume 3881, pages 268–279. Springer Verlag, 2006.
4. A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, A. Ricci, and G. Volpe. Toward real-time multimodal processing: Eyesweb 4.0. In *in Proceedings AISB 2004 Convention: Motion, Emotion and Cognition*, 2004.
5. A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies, Elsevier Science*, 59:213–225, july 2003.
6. A. Camurri, B. Mazzarino, and G. Volpe. Analysis of expressive gesture: The eyesweb expressive gesture processing library. In G.Volpe A. Camurri, editor, *Gesture-based Communication in Human-Computer Interaction*, LNAI 2915. Springer Verlag, 2004.
7. G. Castellano. Human full-body movement and gesture analysis for emotion recognition: a dynamic approach. Paper presented at HUMAINE Crosscurrents meeting, Athens, June 2006.

8. M. DeMeijer. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 28:247 – 268, 1989.
9. A. Drosopoulos, T. Balomenos, S. Ioannou, K. Karpouzis, and S. Kollias. Emotionally-rich man-machine interaction based on gesture analysis. In *Human-Computer Interaction International*, volume 4, page 1372 1376, june 2003.
10. P. E. Gallaher. Individual differences in nonverbal behavior: Dimensions of style. *Journal of Personality and Social Psychology*, 63(1):133–145, 1992.
11. H. Gunes and M. Piccardi. Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. In *Proceedings of Affective Computing and Intelligent Interaction: First International Conference*, October 2005.
12. B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud. Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *Third International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS)*, Utretch, July 2005.
13. B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *The 6th International Workshop on Gesture in Human-Computer Interaction and Simulation*, VALORIA, University of Bretagne Sud, France, 2005.
14. S. Kopp, T. Sowa, and I. Wachsmuth. Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures. In *Gesture Workshop*, pages 436–447, 2003.
15. M. Mancini, R. Bresin, and C. Pelachaud. From acoustic cues to an expressive agent. In *Gesture Workshop*, pages 280–291, 2005.
16. A. Paiva, R. Chaves, M. Piedade, A. Bullock, G. Andersson, and K. Höök. Sentoy: a tangible interface to control the emotions of a synthetic character. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 1088–1089, New York, NY, USA, 2003. ACM Press.
17. C. Pelachaud. Multimodal expressive embodied conversational agents. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 683–689, New York, NY, USA, 2005. ACM Press.
18. B. Reeves and C. Nass. *The media equation: How people treat computers, television and new media like real people and places*. CSLI Publications, Stanford, CA, 1996.
19. D. Reidsma, A. Nijholt, R. Poppe, R. Rienks, and H. Hondorp. Virtual rap dancer: invitation to dance. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pages 263–266, New York, NY, USA, 2006. ACM Press.
20. K. R. Scherer and H. G. Wallbott. Analysis of nonverbal behavior. In *HANDBOOK OF DISCOURSE: ANALYSIS*, volume 2, chapter 11. Academic Press London, 1985.
21. R. Taylor, D. Torres, and P. Boulanger. Using music to interact with a virtual character. In *The 2005 International Conference on New Interfaces for Musical Expression*, 2005.
22. K.R. Thrisson, T. List, C. Pennock, and J. DiPirro. Whiteboards: Scheduling blackboards for interactive robots. In *Twentieth National Conference on Artificial Intelligence: Workshop On Modular Construction of Human-Like Intelligence*, 2005.
23. H. G. Wallbott and K. R. Scherer. Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, 51(4):690–699, 1986.