Multimodal Sensing, Interpretation and Copying of Movements by a Virtual Agent

Elisabetta Bevacqua[2], Amaryllis Raouzaiou[1], Christopher Peters[2], George Caridakis[1], Kostas Karpouzis[1], Catherine Pelachaud[2], and Maurizio Mancini[2]

[1] Image, Video and Multimedia Systems Laboratory, National Technical University of Athens, Greece
http://www.image.ntua.gr
[2] LINC, IUT de Montreuil, Université de Paris 8
http://www.univ-paris8.fr

Abstract. We present 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. Thus, the copied behaviour may be altered and need not share all of the characteristics of the original made by the human.

1 Introduction

The ability for an agent to provide feedback to a user is an important means for signalling to the world that they are animate, engaged and interested. Feedback influences the plausibility of an agent's behaviour with respect to a human viewer and enhances the communicative experience.

In this paper, we present 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 [1]. Thus, the copied behaviour may be altered and need not share all of the characteristics of the original made by the human.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 164–174, 2006.
© Springer-Verlag Berlin Heidelberg 2006

Of particular interest is that a subset of the framework has already been used in conjunction with synthetic vision to implement conversation initiation behaviours between agents in a virtual environment based on their goals and perception of the attention shown by others in them through gaze [2]. In this paper, we also describe how the same framework may be used with real world input. This is an important feature of our work, as one of our long term objectives is to endeavor towards an agent framework that allows agents to interact in a seamless manner with both real and virtual environments in order to further investigate the inherent interactional differences between such environments.

2 State of the Art

There is a long history of interest in the problem of recognising emotion from facial expressions, and extensive studies on face perception during the last twenty years [3]. Ekman and Friesen elaborated a scheme to annotate facial expressions named Facial Action Coding System (FACS) [4] to manually describe facial expressions, using still images of, usually extreme, facial expressions. In the nineties, automatic facial expression analysis research gained much interest mainly thanks to progress in the related fields such as image processing (face detection, tracking and recognition) and the increasing availability of relatively cheap computational power. Head pose and especially facial expression having a very important role in the Human Computer Interaction (HCI), many researchers tackle the problem of facial expression analysis [5], [6] and head movements [7], [8]. Regarding feature-based techniques, Donato et al [9] tested different features for recognizing facial Action Units (AUs) and inferring the facial expression in the frame. Analysis of the emotional expression of a human face requires a number of pre-processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face.

The detection and interpretation of hand gestures has become an important part of HCI in recent years [10]. The HCI interpretation of gestures requires that dynamic and/or static configurations of the human hand, arm, and even other parts of the human body, be measurable by the machine. First attempts to address this problem resulted in mechanical devices that directly measure hand and/or arm joint angles and spatial position. The so-called *glove-based* devices best represent this type of approach. Analysing hand gestures is a comprehensive task involving motion modeling, motion analysis, pattern recognition, machine learning, and even psycholinguistic studies. The first phase of the recognition task is choosing a model of the gesture. Among the important problems involved in the analysis are those of hand localization, hand tracking, and selection of suitable image features. The computation of model parameters is followed by gesture recognition.Hand localization is locating hand regions in image

sequences. Skin color offers an effective and efficient way to fulfill this goal. An interesting approach of gesture analysis research [11] treats a hand gesture as a two- or three dimensional signal that is communicated via hand movement from the part of the user; as a result, the whole analysis process merely tries to locate and track that movement, so as to recreate it on an avatar or translate it to specific, predefined input interface, e.g. raising hands to draw attention or indicate presence in a virtual classroom. There are many systems available for synthesising the animation of a virtual agent. Badler's research group developed EMOTE (Expressive MOTion Engine [12]), a parameterized model that procedurally modifies the affective quality of 3D character's gestures and postures motion. From EMOTE the same research group derived FacEMOTE [13], a method for facial animation synthesis that altered pre-computed expressions by setting a small set of high level parameters taken from Laban Parameters. Wachsmuth's group [14] described a virtual agent capable of imitating natural gestures performed by a human using captured data. Imitation is conducted on two levels: when mimicking, the agent extracts and reproduces the essential form features of the stroke which is the most important gesture phase; the second level is a meaning-based imitation level that extracts the semantic content of gestures in order to re-express them with different movements.

3 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.

Visual input is processed by computer vision [15] (see Section 4.1) or synthetic vision techniques [2], as appropriate, and stored in a short-term sensory storage. 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. FAPs $\rightarrow$ 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 $\rightarrow$ another set of FAPs). The generation module [16, 17] synthesises the final desired agent behaviours (Section 4.2). In this paper we present a system of an ECA able to perceive facial and gesture expressions performed by a real user. In order to demonstrate such a capability we present, in Section 5, a simple scenario where the ECA perceives and reproduces the user's movements by using a generation module. That is, in our system, the resulting animation is not a pure copy of the perceived data. In the future, we aim to use this capability to implement a more complex decisional model: by *decisional model*, we refer to a model capable of deciding which movement the ECA will perform, in accordance with the current user's behaviour.

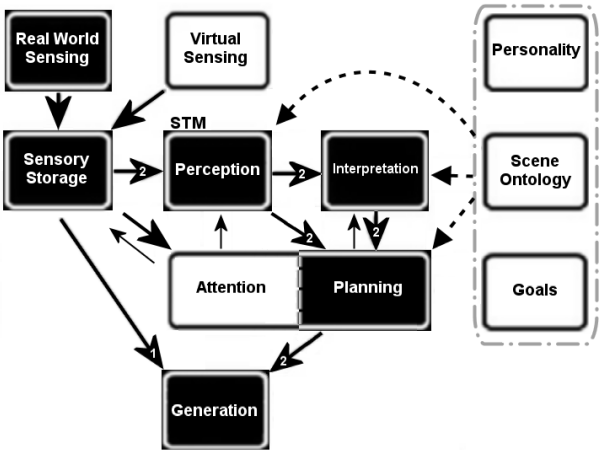

Fig. 1. The general framework that embeds the current scenario. Large arrows indicate the direction of information flow, small arrows denote control signals, while arrows with dashed lines denote information availability from modules associated with long term memory. Modules with a white background are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

4 Description of Expressivity Model

The expressivity of behaviors, that is the way behaviors are executed, is an integral part of the communication process. Several researchers ([18], [19], [20], [21], [22]) have investigated human motion characteristics and encoded them into dimensions. In particular Wallbott and Sherer have conducted perceptual studies that show that human beings are able to perceive and recognise a set of these dimensions [19].

Some authors refer to body motion using dual categories such as slow/fast, small/expansive, weak/energetic, small/large, unpleasant/pleasant. To model expressivity, in our framework we use 6 parameters, derived from perceptual studies conducted by [19], each represented by dual category:

- *Overall Activation*: quantity of movements in a timespan
- *Spatial Extent*: amplitude of movements (e.g., amount of space taken up by body or of emotion arousal)
- *Temporal Extent*: duration of movements (e.g., quick vs sustained actions)
- *Fluidity*: smoothness and continuity of overall movement (e.g., smooth vs jerky)
- *Power*: dynamic properties of the movement (e.g., weak vs strong)
- *Repetition*: tendency to rhythmic repeats of specific movement.

Evaluation studies conducted on our model show that spatial and temporal dimensions are easily recognised, whereas fluidity and power are more difficult to interpret. Repetition of a gesture has been often mistaken as being a single complex gesture rather than the repetition of a simple gesture [23].

4.1 Analysis

Facial analysis includes a number of processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face. Although FAPs [24] provide all the necessary elements for MPEG-4 compatible animation, we cannot use them for the analysis of expressions from video scenes, due to the absence of a clear quantitative definition framework. In order to measure FAPs in real image sequences, we have to define a mapping between them and the movement of specific feature points (FPs), which correspond to salient points on the human face [25]. The proposed facial analysis subsystem can detect facial expressions in good lighting conditions. Additionally, the face should not be in an angle omitting characteristic facial features like eye or lip corner. The proposed facial feature extraction scheme is based on a hierarchical, robust scheme, where soft a priori assumptions are made on the pose of the face or the general location of the features in it. Gradual revelation of information concerning the face is supported under the scope of optimisation in each step of the hierarchical scheme, producing a posteriori knowledge about it and leading to a step-by-step visualisation of the features in search. Face detection is performed first through detection of skin segments or blobs, merging them based on the probability of their belonging to a facial area, and identification of the most salient skin color blob or segment. Primary facial features, such as eyes, mouth and nose, are treated as major discontinuities on the segmented, arbitrarily rotated face. Following face detection, morphological operations, erosions and dilations, taking into account symmetries, are used to define the most probable blobs within the facial area to include the eyes and the mouth. Searching through gradient filters over the eyes and between the eyes and mouth provide estimates of the eyebrow and nose positions. Based on the detected facial feature positions, feature points are computed and evaluated.

The next step of the system is the tracking of head and hand. The input image sequences of the gesture analysis subsystem are real videos captured at an acted session. The gestures that our subsystem can detect should be distinguishable in the 2-D frame we have at our disposal, e.g. a hand moving towards the camera-in the vertical plane cannot be detected. Several approaches have been reviewed for a gesture analysis module. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin detection and tracking module. The general process involves the creation of moving skin masks, namely skin color areas that are tracked between subsequent frames. By tracking the centroid of those skin masks, an estimate of the user's movements is produced. A priori knowledge concerning the human body and the circumstances when filming the gestures was incorporated into the module indicating the different body parts (head, right hand, left hand). For each frame (Figure 2, top left) a skin color probability matrix is computed by calculating the joint probability of the Cr/Cb image values (Figure 2, top center). The skin color mask is then obtained from the skin probability matrix using

thresholding (Figure 2, top right). Possible moving areas are found by thresholding the pixels difference between the current frame and the next, resulting in the possible-motion mask (Figure 2, bottom center). This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color and motion masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. All described morphological operations are carried out with a disk-structuring element with a radius of 1% of the image width. The distance transform of the color mask is first calculated (Figure 2, bottom right) and only objects above the desired size are retained. These objects are used as markers for the morphological reconstruction of the initial color mask. The color mask is then closed to provide better centroid calculation. For the next frame, a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area. In the case of hand object merging and splitting, e.g., in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand. The tracking algorithm is responsible for classifying the skin regions in the image sequence of the examined gesture based on the skin regions extracted from the described method.

As far as expressivity dimensions are concerned, they have been designed for communicative behaviours only [17], [26]. Each dimension acts differently for

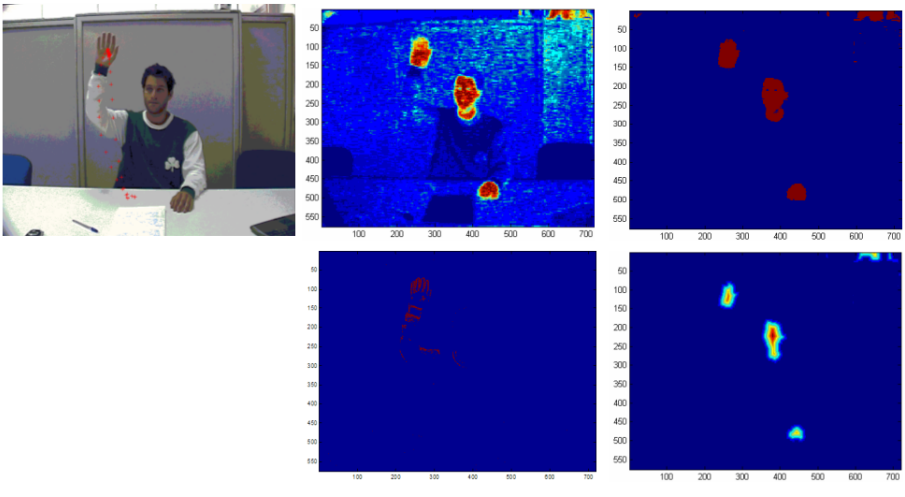

Fig. 2. Top left: example of video frame. Top center: Cr/Cb image. Top right: skin color mask. Bottom center: possible-motion mask. Bottom right: distance transform of the color mask.

each modality. For an arm gesture, expressivity works at the level of the phases of the gesture: for example the preparation phase, the stroke, the hold as well as on the way two gestures are co-articulated [27, 20]. We consider the six dimensions of expressivity as defined in Section 4. Overall activation is considered as the quantity of movement during a conversational turn. In our case it is computed as the sum of the motion vectors' norm (as shown in formula 1). Spatial extent is modeled by expanding or condensing the entire space in front of the human that is used for gesturing and is calculated as the maximum Euclidean distance of the position of the two hands (see formula 2).

$$OA = \sum_{i=0}^{n} \mid \boldsymbol{r}(i) \mid + \mid \boldsymbol{l}(i) \mid . \tag{1}$$

$$SE = \max(\mid d(r(i) - l(i)) \mid). \tag{2}$$

The average spatial extent is also calculated for normalisation reasons. The temporal parameter of the gesture determines the speed of the arm movement of a gesture's meaning carrying stroke phase and also signifies the duration of movements (e.g., quick versus sustained actions). Fluidity differentiates smooth/graceful from sudden/jerky ones. This concept seeks to capture the continuity between movements, as such, it seems appropriate to modify the continuity of the arms' trajectory paths as well as the acceleration and deceleration of the limbs. To extract this feature from the input image sequences we calculate the sum of the variance of the norms of the motion vectors. Finally, the power is identical to the first derivative of the motion vectors calculated in the first steps.

4.2 Synthesis

Table 1 shows the effect that each expressivity parameter has on the production of head movements, facial espressions and gestures. The *Spatial Extent* (SPC) parameter modulates the amplitude of the movement of arms, wrists (involved in the animation of a gesture), head and eyebrows (involved in the animation of a facial expression); it influences how wide or narrow their displacement will be. For example let us consider the eyebrows raising in the expression of surprise: if the value of the *Spatial Extent* parameter is very high the final position of the eyebrows will be very high (i.e. the eyebrows move under a strong of muscular contraction). The *Temporal Extent* (TMP) parameter shortens or lengthens the motion of the preparation and retraction phases of the gesture as well as the onset and offset duration for facial expression. It speeds up or slows down the rising/lowering of the eyebrows. The animation of the agent is generated by defining *key frames* and computing the interpolation curves passing through these frames using TCB-Splines. The *Fluidity* (FLT) and *Power* (PWR) parameters act on the interpolation curves. *Fluidity* increases/reduces the continuity of the curves allowing the system to generate more of less smooth animations. Let us consider its effect on the head: if the value of the *Fluidity* parameter is very low the resulting curve of the head movement will appear as generated through linear interpolation resulting as jerky movement. *Power* introduces a

gesture/expression overshooting, that is a little lapse of time in which the body part involved by the gesture reaches a point in space further than the final one. For example the frown displayed in the expression of anger will be stronger for a short period of time, and then the eyebrows will backtrack to reach the final position. The last parameter, *Repetition* (REP) increases the number of stroke of gestures to obtain repetition of the gestures themselves in the final animation. Let us consider the gesture "wrists going up and down in front of the body with open hands and palms up", a high value of the *Repetition* parameter will increase the number of the up and down movements. On the other hand this parameter decreases the time period of head nods and head shakes to obtain more nods and shakes in the same lapse of time.

The synthesis module is MPEG-4 compatible. Facial expressions are described by FAPs value. Gestures are computed through the interpolation of a sequence of static positions defined by shoulder and arm rotation (arm position), hand shape (chosen in a set of predefined shapes) and palm orientation [28].

Table 1. Effects of Expressivity parameters over head, facial expression and gesture

	HEAD	FACIAL EXPRESSION	GESTURE
SPC	wider/narrower movement	increased/decreased muscular contraction	wider/narrower movement
TMP	shorter/longer movement speed	shorter/longer onset and offset	shorter/longer speed of preparation and retraction phases
FLT	increases/reduces continuity of head movement	increases/reduces continuity of muscular contraction	increases/reduces continuity between consecutive gestures
PWR	higher/shorter head overshooting	higher/shorter muscular contraction overshooting	more/less stroke acceleration
REP	more/less number of nods and shakes	not implemented yet	more/less number of repetitions of the same stroke

5 Application Scenario and Implementation

In section 3 we described the general framework of the system we are aiming at that is able to analyse a real scene and generate the animation of a virtual agent. In this section we present a scenario that is a partial implementation of this framework. 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 intrinsic expressivity parameters extracted from the actor's performance. Figure 3 is a sub-set of the architecture shown in Figure 1. It describes the architecture required by our current application scenario. The modules of this diagram correspond to the modules in black

in the original architecture. 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 (see Section 4) 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, consisting of variation of FAPs and BAPs values during time, is calculated through the Face and the

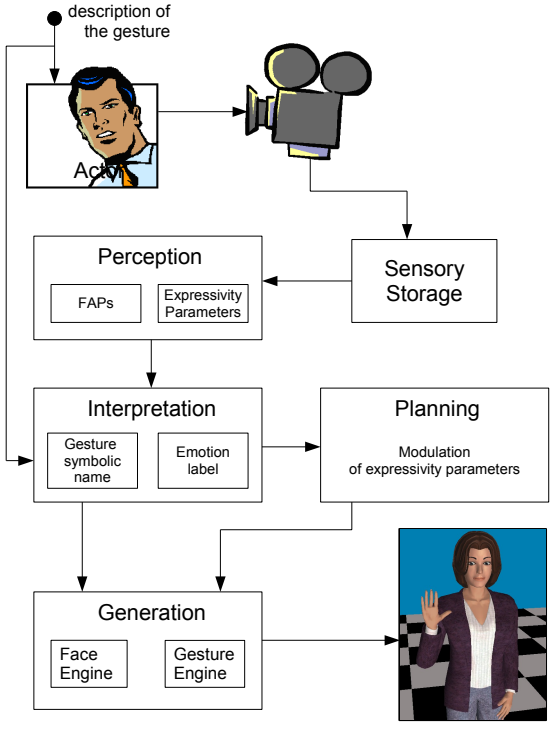

Fig. 3. Diagram of the proposed implementation

Gesture Engine and displayed by the virtual agent. The system does not work in real-time yet, but we aim to develop real-time capabilities in the near future. We also intend to evaluate our system through perceptual tests in order to estimate the correctness of movements.

6 Conclusions

We have presented our general framework consisting of a number of interconnected modules and a sample scenario whereby an agent senses, interprets and copies a range of facial and gesture expression from a person in the real-world. The animation of the agent is based on different types of data: raw parameters values, emotion labels, expressivity parameters, and symbolic gestures specification. To do so the system is able to perceive and interpret gestural and facial expressions made by an actor.

A very interesting extension to the framework would be the addition of visual attention capabilities. As seen in the design of Figure 1, attention may be used to select certain information in the sensory storage, perception or interpretation stages for access to further stages of processing, as well as modulating planning and for some behaviour generation, such as orienting an agent's gaze. An attention system, applicable to both real and virtual environments, in a unified framework, is an interesting prospect. Finally, we also aim to use the analysis-synthesis loop as a learning phase to refine the synthesis model of expressivity and of behaviour.

References

1. Martin, J.C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., Pelachaud, C.: Levels of representation in the annotation of emotion for the specification of expressivity in ecas. In: International Working Conference on Intelligent Virtual Agents, Kos, Greece (2005) 405–417
2. Peters, C.: Direction of attention perception for conversation initiation in virtual environments. In: International Working Conference on Intelligent Virtual Agents, Kos, Greece (2005) 215–228
3. Scherer, K., Ekman, P.: Approaches to Emotion. Lawrence Erlbaum Associates (1984)
4. Ekman, P., Friesen, W.: The Facial Action Coding System, Consulting Psychologists Press. San Francisco, CA (1978)
5. Tian, Y., Kanade, T., Cohn, J.: Facial expression analysis. In: Handbook of face recognition, S.Z. Li and A.K. Jain, ed.,. (2003)
6. Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: Computer Vision and Pattern Recognition. Volume 2., Computer Vision and Pattern Recognition, CVPR 2005 (2005) 568–573
7. Morency, L.P., Sidner, C., Lee, C., Darrell, T.: Contextual recognition of head gestures. In: Proceedings of ICMI'05, Trento, Italy (2005)
8. Gratch, J., Marsella, S., Maatman, M., Gratch, J., Marsella, S.: Natural behavior of a listening agent. In: Proceedings of IVA05, Kos, Greece (2005)

9. Donato, G., Bartlett, M., Hager, J., Ekman, P., Sejnowski, T.: Classifying facial actions. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 21. (1999)
10. Wu, Y., Huang, T.: Hand modeling, analysis, and recognition for vision-based human computer interaction. In: IEEE Signal Processing Magazine. Volume 18. (2001) 51–60
11. Wexelblat, A.: An approach to natural gesture in virtual environments. In: ACM Transactions on Computer-Human Interaction. Volume 2. (1995) 179–200
12. Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. In: ACM SIGGRAPH '00, New Orleans, LA (2000) 173–182
13. Byun, M., Badler, N.: FacEMOTE: Qualitative parametric modifiers for facial animations. In: Symposium on Computer Animation, San Antonio, TX (2002)
14. Kopp, S., Sowa, T., Wachsmuth, I.: Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures. In: Gesture Workshop. (2003) 436–447
15. Rapantzikos, K., Avrithis, Y.: An enhanced spatiotemporal visual attention model for sports video analysis. In: International Workshop on content-based Multimedia indexing (CBMI), Riga, Latvia (2005)
16. Pelachaud, C., Bilvi, M.: Computational model of believable conversational agents. In Huget, M.P., ed.: Communication in Multiagent Systems. Volume 2650 of Lecture Notes in Computer Science. Springer-Verlag (2003) 300–317
17. Hartmann, B., Mancini, M., Pelachaud, C.: Implementing expressive gesture synthesis for embodied conversational agents. In: Gesture Workshop, Vannes (2005)
18. Johansson, G.: Visual perception of biological motion adn a model for its analysis. Perception and Psychophysics **14** (1973) 201–211
19. Wallbott, H.G., Scherer, K.R.: Cues and channels in emotion recognition. Journal of Personality and Social Psychology **51** (1986) 690–699
20. Gallaher, P.E.: Individual differences in nonverbal behavior: Dimensions of style. Journal of Personality and Social Psychology **63** (1992) 133–145
21. Ball, G., Breese, J.: Emotion and personality in a conversational agent. In J. Cassell, J. Sullivan, S.P., Churchill, E., eds.: Embodied Conversational Characters. MITpress, Cambridge, MA (2000)
22. Pollick, F.E.: The features people use to recognize human movement style. In Camurri, A., Volpe, G., eds.: Gesture-Based Communication in Human-Computer Interaction - GW 2003. Number 2915 in LNAI. Springer (2004) 10–19
23. Hartmann, B., Mancini, M., Buisine, S., Pelachaud, C.: Design and evaluation of expressive gesture synthesis for embodied conversational agents. In: Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Utretch (2005)
24. Tekalp, A., Ostermann, J.: Face and 2-d mesh animation in MPEG-4. In: Signal Processing: Image Communication. Volume 15. (2000) 387–421
25. Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., Kollias, S.: Parameterized facial expression synthesis based on mpeg-4. In: EURASIP Journal on Applied Signal Processing. Volume 2002., Hindawi Publishing Corporation (2002) 1021–1038
26. Wallbott, H.G.: Bodily expression of emotion. In: European Journal of Social Psychology. Volume 28. (1998) 879–896
27. Harrigan, J.A.: Listener's body movements and speaking turns. In: Communication Research. Volume 12. (1985) 233–250
28. Hartmann, B., Mancini, M., Pelachaud, C.: Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. In: Computer Animation'02, Geneva, Switzerland, IEEE Computer Society Press (2002)