Lang Resources & Evaluation (2007) 41:367–388
DOI 10.1007/s10579-007-9057-1

Virtual agent multimodal mimicry of humans

George Caridakis · Amaryllis Raouzaiou · Elisabetta Bevacqua ·
Maurizio Mancini · Kostas Karpouzis · Lori Malatesta · Catherine Pelachaud

Published online: 10 January 2008

Abstract This work is about multimodal and expressive synthesis on virtual agents, based on the analysis of actions performed by human users. As input we consider the image sequence of the recorded human behavior. Computer vision and image processing techniques are incorporated in order to detect cues needed for expressivity features extraction. The multimodality of the approach lies in the fact that both facial and gestural aspects of the user's behavior are analyzed and processed. The mimicry consists of perception, interpretation, planning and animation of the expressions shown by the human, resulting not in an exact duplicate rather than an expressive model of the user's original behavior.

G. Caridakis (✉) · A. Raouzaiou · K. Karpouzis · L. Malatesta
Image, Video and Multimedia Systems Laboratory, National Technical University of Athens,
Athens, Greece
e-mail: gcari@image.ece.ntua.gr
URL: http://www.image.ntua.gr

A. Raouzaiou
e-mail: araouz@image.ece.ntua.gr

K. Karpouzis
e-mail: kkarpou@image.ece.ntua.gr

L. Malatesta
e-mail: lori@image.ece.ntua.gr

E. Bevacqua · M. Mancini · C. Pelachaud
LINC, IUT de Montreuil, Universite de Paris 8, Paris, France
URL: http://www.univ-paris8.fr

E. Bevacqua
e-mail: e.bevacqua@iut.univ-paris8.fr

M. Mancini
e-mail: m.mancini@iut.univ-paris8.fr

C. Pelachaud
e-mail: pelachaud@iut.univ-paris8.fr

Keywords Emotional ECA synthesis · Expressivity features · Facial features extraction · Gesture analysis · Virtual agents

1 Introduction

The ability of life-like virtual agents to provide expressive feedback to a user is an important aspect to support their naturalness. Both analysis and synthesis of multimodal cues constitute an important part of human–computer interaction (HCI). Multimodal feedback influences the plausibility of an agent's behavior with respect to a human viewer and enhances the communicative experience. As a general rule, mimicry is an integral, while often unaware, part of human–human interaction (Lakin et al. 2003; Chartland et al. 2005). In this framework, a "loop" can be defined, where attachment results in non-conscious imitation of the other party's body posture, hand gestures and even facial expressions, which in turn improves one's relationship with others (van Swol 2003). Extending this paradigm to human–computer interaction, one is safe to expect that users interacting via an interface which provides system prompts via an affective Embodied Conversational Agent (ECA) and receives user input via natural means of communication (facial expressions, hand, head, and body gestures) feel more comfortable than in the case of interacting via the usual "windows, mouse, pointer" archetype (Oviatt 1999).

While affective arousal modulates all human communicative signals (Ekman and Friesen 1969), the visual channel facial expressions and body/hand gestures) is deemed to be the most important in the human judgment of behavioral cues (Ambady et al. 1992), since human observers seem to be most accurate in their judgment when looking at the face and the body than depending on the voice alone. This fact indicates that people rely on shown facial expressions to interpret someone's behavioral disposition and to a lesser degree on shown vocal expressions. However, although basic researchers have been unable to identify a set of voice cues that reliably discriminate among emotions, listeners seem to be accurate in decoding emotions from voice cues (Juslin and Scherer 2005). Thus, human affect analysis corpora should at least include facial expressions and related features as a modality and preferably they should also cater for perceiving either body gestures or speech prosody (Caridakis et al. 2006). Finally, while too much information from different channels seem to be confusing to human judges (Oviatt 1999), resulting in less accurate judgments of shown behavior when more observation channels are available (e.g., face, body, and speech), combining those multiple modalities (including speech and physiology) may prove appropriate for realization of automatic human affect analysis.

Our work concentrates on the intermediate procedures needed for an agent to properly sense, interpret and copy a range of facial and gesture expression from a person in the real-world as can be seen on Fig. 1. Image sequences are processed as to extract prominent facial points (Facial Definition Parameters—FDPs), their deformation (Facial Animation Parameters—FAPs) and the position of the user's hand (Raouzaiou et al. 2002). FDPs and FAPs are defined in the framework of

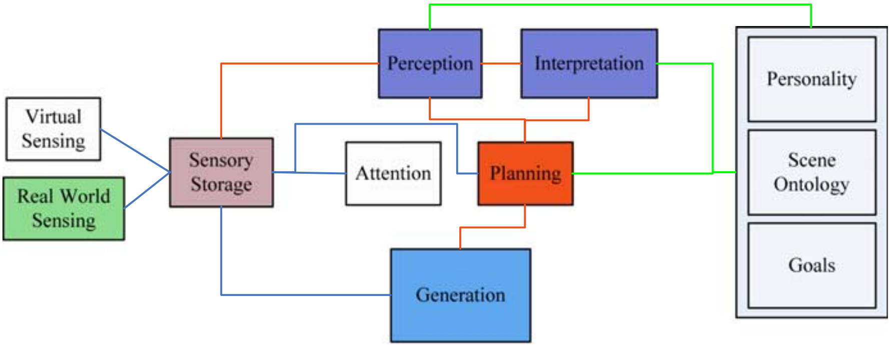

Fig. 1 An abstract overview of our approach. Modules in color have been also implemented in the scenario described in Sect. 5

MPEG-4 standard and provide a standardized means of modeling facial geometry and expressivity and are strongly influenced from the Action Units (AUs) defined in neurophysiological and psychological studies (Ekman and Friesen 1978). The adoption of token-based animation in the MPEG-4 framework benefits the definition of emotional states, since the extraction of simple, symbolic parameters is more appropriate to analyze, as well as synthesize facial expression and hand gestures. This information is used both to calculate expressivity features and to model and classify the emotional aspects of a user's state. It is then processed further in a framework for agent perception, planning, and behavior generation in order to perceive, interpret and copy a number of gestures and facial expressions corresponding to those made by the human; related theoretical aspects are introduced in Sect. 2. By *perceive*, we mean that the copied behavior may not be an exact duplicate of the behavior made by the human and sensed by the agent, but may rather be based on some level of interpretation of the behavior (Martin et al. 2005); the ability of our ECA to perceive facial and hand expressivity in presented in Sect. 3.

Elaboration of this data involves a symbolic and semantic processing, high-level representation and long-term planning processes. Moreover it implies an interpretation of the viewed expression (e.g., FAPs to 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 to another set of FAPs). The generation module (Pelachaud and Bilvi 2003; Hartmann et al. 2005a), which synthesizes the final desired agent behaviors is described in Sect. 4, while the capability of our ECA to sense facial and gesture expressions performed by a real user is illustrated in Sect. 5 by means of a simple scenario where the ECA perceives and reproduces the user's movements. Such perception makes sure that the resulting animation is not a pure copy. In the future, we aim to use this capability to implement a more complex decisional model, which caters for deciding which movement the ECA will perform, according also to the current user's behavior and evaluating the soundness of the proposed approach using the scheme discussed in Sect. 6.

2 Background

2.1 Fusion of modalities for analysis and synthesis

Although there is a satisfactory amount of literature on many modules of the proposed architecture, the work related with the fusion of different modalities as well as the association of analysis and synthesis is very scarce (Peters 2005). There is a long history of interest in the problem of recognizing emotion from facial expressions, and extensive studies on face perception during the last 20 years (Scherer and Ekman 1984). Ekman and Friesen (1978) elaborated a scheme to annotate facial expressions named Facial Action Coding System (FACS) 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. Regarding feature-based techniques, Donato et al. (1999) tested different features for recognizing facial 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 (Wu and Huang 2001). 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 solutions' group. Analyzing 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 (Wexelblat 1995) 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 for animation synthesis of a virtual agent. Badler's research group developed EMOTE (Expressive MOTion Engine (Chi et al. 2000)), 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 (Byun and Badler 2002), a method for facial animation synthesis varying pre-existent expressions by setting a small set of high level parameters. Wachsmuth's group (Kopp et al. 2003) described a virtual agent able to imitate natural gestures performed by a human using captured data. The imitation is done on two levels: the first one is the mimicking level, the agent extracts and reproduces the essential form features of the stroke which is the most important gesture phase; the second level is the meaning-based imitation level that extracts the semantic content of gestures to re-express them with different movements.

2.2 Emotion representation

Psychologists have examined a broader set of emotions, but very few of the studies provide results which can be exploited in computer graphics and machine vision fields. One of these studies, carried out by Whissel (1989), suggests that emotions are points in a space spanning a relatively small number of dimensions, which seem to occupy two axes: *activation* and *evaluation*. The activation-evaluation space is a representation that is both simple and capable of capturing a wide range of significant issues in emotion. It rests on a simplified treatment of two key themes (Fig. 2):

- *Valence* (Evaluation level): The clearest common element of emotional states is that the person is materially influenced by feelings that are "valenced," i.e., they are centrally concerned with positive or negative evaluations of people or things or events. The link between emotion and valencing is widely agreed (horizontal axis).
- *Activation* level: Research has recognized that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated

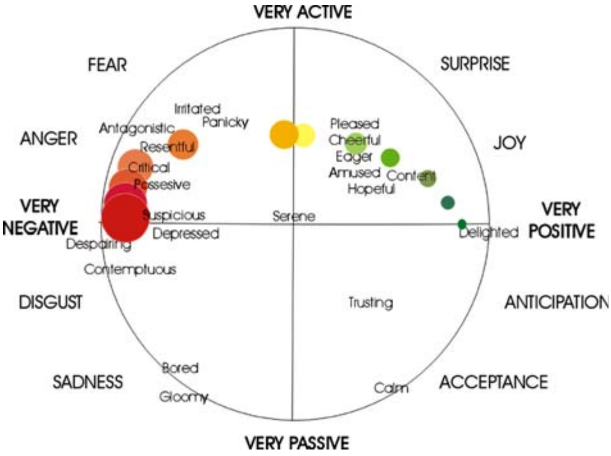

Fig. 2 The activation–evaluation space

activation level, i.e., the strength of the person's disposition to take some action rather than none (vertical axis).

3 Analysis

3.1 System overview

Expressive features, useful during the synthesis module, are based upon the extraction of facial and gestural information. For the face, FAP (Tekalp and Ostermann 2000) values are extracted using the methodology depicted in Fig. 3 and described in Sect. 3.2. (Ioannou et al. 2005). As for gesture analysis, the hand and head relative distances normalized wrt the head size are used. Figure 4 is indicative of the procedure for hand detection and tracking.

In order to have both the required image resolution to extract facial features and satisfy the spatial requirements of gesture processing two individual video streams were acquired from different cameras. We have chosen such a setup because the resolution required for facial features extraction is much larger than the one for hand gestures tracking. This could only be achieved if one camera zoomed in the subject's face. Video streams were synchronized manually prior to the processing step.

3.2 Facial feature extraction

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 provide all the necessary elements for MPEG-4 compatible

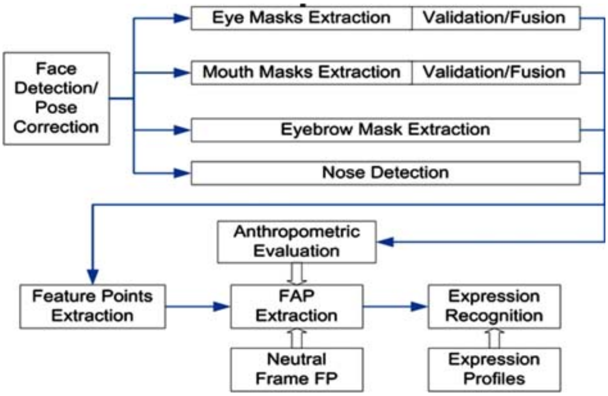

Fig. 3 Diagram of the FAP extraction methodology and the expression recognition module

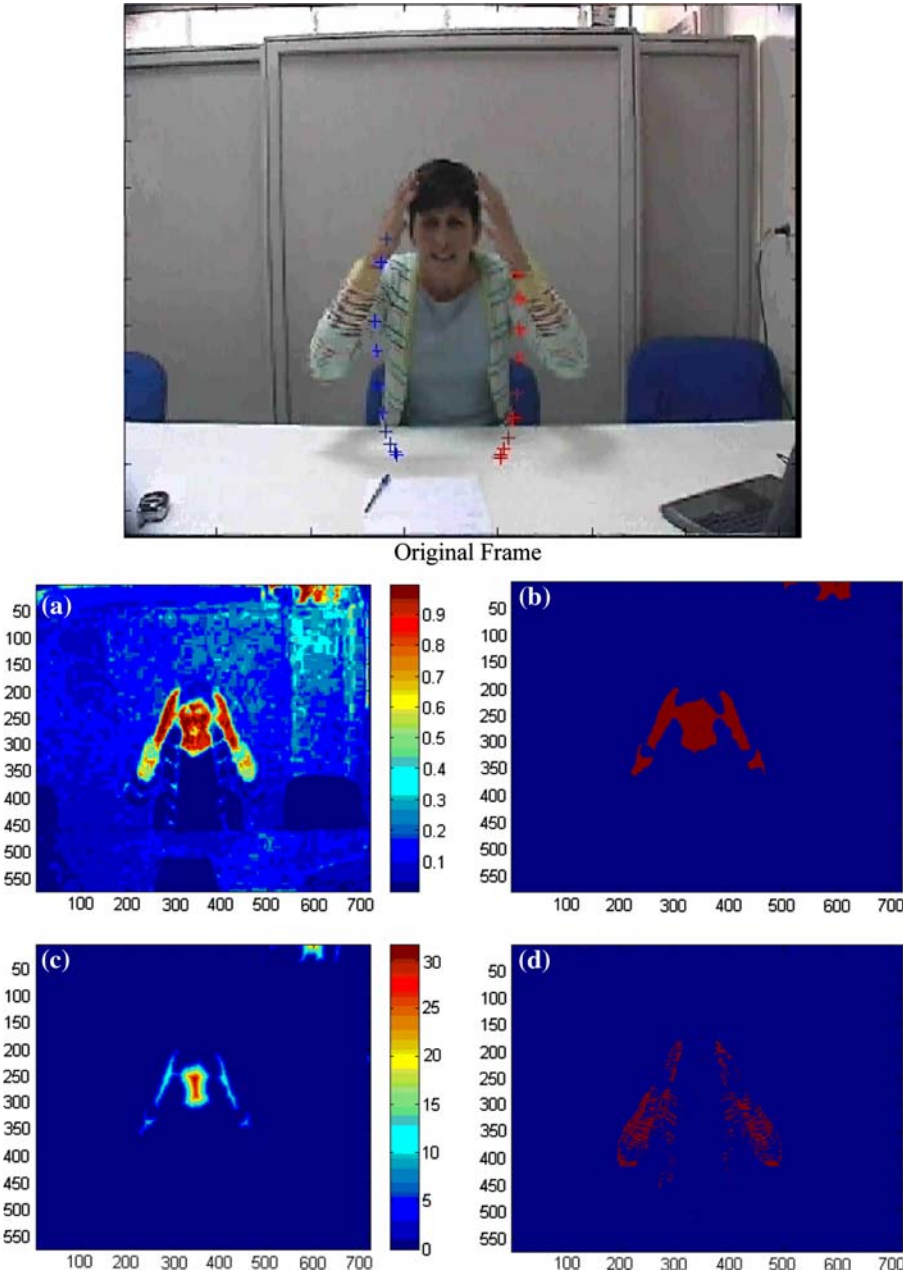

Fig. 4 Key steps in hand detection and tracking (**a**) skin probability (**b**) thresholding & morphology operators (**c**) distance transformation, and (**d**) frame difference

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 facial definition parameters (FDPs) feature points (FPs), which correspond to salient points on the human face.

The face is first located, so that approximate facial feature locations can be estimated from the head position and rotation. Face roll rotation is estimated and corrected and the head is segmented focusing on the following facial areas: left eye/eyebrow, right eye/eyebrow, nose, and mouth. Each of those areas, called feature-candidate areas, contains the features whose boundaries need to be extracted for our purposes. Inside the corresponding feature-candidate areas precise feature extraction is performed for each facial feature, i.e., eyes (Fig. 5), eyebrows (Fig. 6), mouth and nose, using a multi-cue approach, generating a small number of intermediate feature masks. Feature masks generated for each facial feature are fused together to produce the final mask for that feature. The mask fusion process uses anthropometric criteria to perform validation and weight assignment on each intermediate mask; each feature's weighted masks are then fused to produce a final mask along with confidence level estimation. The edges of the final masks are considered to be the extracted FPs as can be seen in Fig. 7.

Measurement of FAPs requires the availability of a frame where the subject's expression is found to be neutral. This frame will be called the neutral frame and is manually selected from video sequences to be analyzed or interactively provided to the system when initially brought into a specific user's ownership. The final feature masks are used to extract 19 FPs; FPs obtained from each frame are compared to FPs obtained from the neutral frame to estimate facial deformations and produce the FAPs. Confidence levels on FAP estimation are derived from the equivalent feature

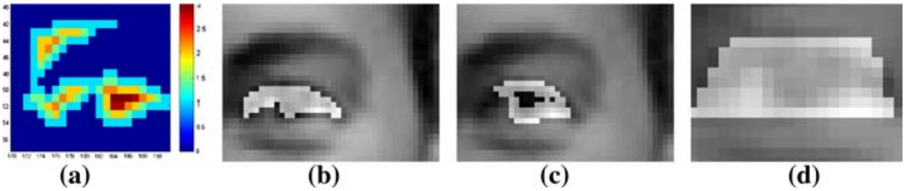

Fig. 5 Eye masks

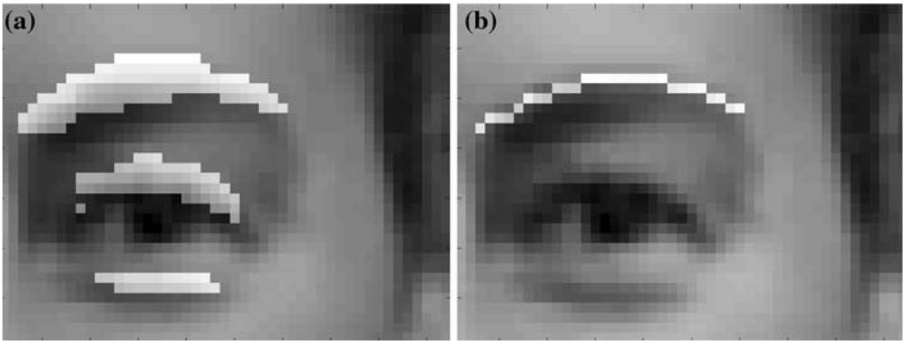

Fig. 6 (**a**) Eyebrow-candidates and (**b**) selected eyebrow mask

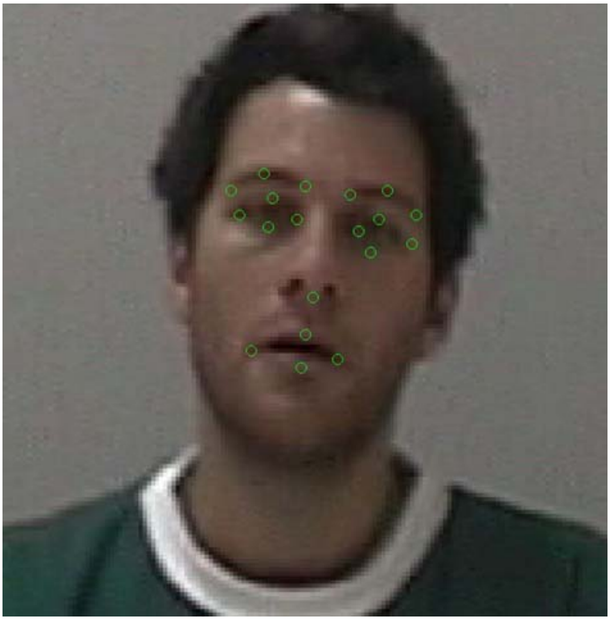

Fig. 7 Facial FPs on one of the subjects

point confidence levels. The FAPs are used along with their confidence levels to provide the facial expression estimation.

Concerning the robustness of the proposed algorithm we had a subset of the available dataset manually annotated by a number of reviewers and compared this annotation versus the automatic one obtained by the described system. As a metric for robustness evaluation we incorporated a modified Williams' Index (Williams 1976) where the results of automatic feature extraction are considered as one of the reviewers. When WI is larger than 1, the computer generated mask disagrees less with the observers than the observers disagree with each other. Figures 8 and 9 show the distribution of WI for eyes/mouth and eyebrows regions, respectively. These results indicate that the described algorithm is efficient given the image is of acceptable quality, the head pose is quite frontal so that feature occlusion does not occur.

3.3 Hand detection and tracking

Regarding gesture analysis, several approaches have been reviewed for the head-hand tracking module all of them mentioned both in Wu and Huang (1999) and in Ong and Ranganath (2005). From these only video based methods were considered since motion capture or other intrusive techniques would interfere with the person's emotional state. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin detection and tracking module.

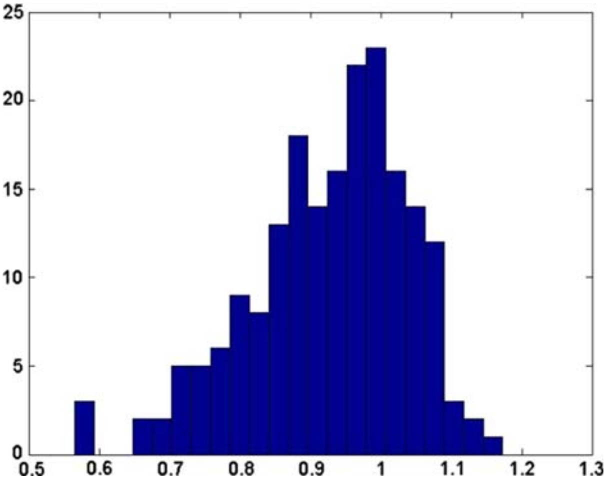

Fig. 8 Williams index distribution (average on eyes and mouth)

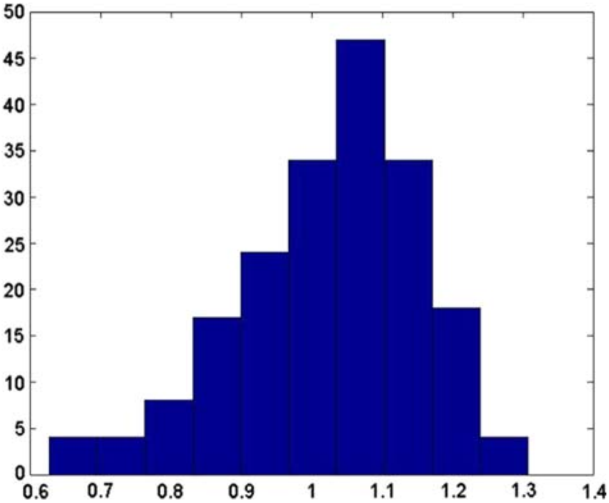

Fig. 9 Williams index distribution (average on left and right eyebrows)

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, we produce an estimate of the user's movements. 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 a skin color probability matrix is computed by calculating the joint probability of the Cr/Cb image values. The skin color mask is then obtained from the skin probability matrix using thresholding. Possible moving areas are found by thresholding the pixels' difference between the current frame and the next, resulting in the possible-motion mask. 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 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 Sagittal plane information of the gesture was ignored since it would require depth information from the video stream and it would make the performance of the proposed algorithm very poor or would require a side camera and parallel processing of the two streams. The described algorithm is lightweight, allowing a rate of around 12 fps on a usual PC during our experiments, which is enough for continuous gesture tracking. The object correspondence heuristic makes it possible to individually track the hand segments correctly, at least during usual meaningful gesture sequences. In addition, the fusion of color and motion information eliminates any background noise or artifacts, thus reinforcing the robustness of the proposed approach.

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. Skin region size, distance with reference to the previous classified position of the region, flow alignment and spatial constraints. These criteria ensure that the next region selected to replace the current one is approximately the same size, close to the last position and moves along the same direction as the previous one as long as the instantaneous speed is above a certain threshold. As a result each candidate region is being awarded a bonus for satisfying these criteria or is being penalized for failing to comply with the restrictions applied. The winner region is appointed as the reference region for the next frame. The criteria don't have an eliminating effect, meaning that if a region fails to satisfy one of them is not being excluded from the process, and the bonus or penalty given to the region is relative to the score achieved in every criterion test. The finally selected region's score is thresholded so that poor scoring winning regions are excluded. In this case the position of the body part is unchanged with reference to that in the previous frame. This feature is especially useful in occlusion cases when the position of the body part remains the same as just before occlusion occurs. After a certain number of frames the whole process is reinitialized so that a possible

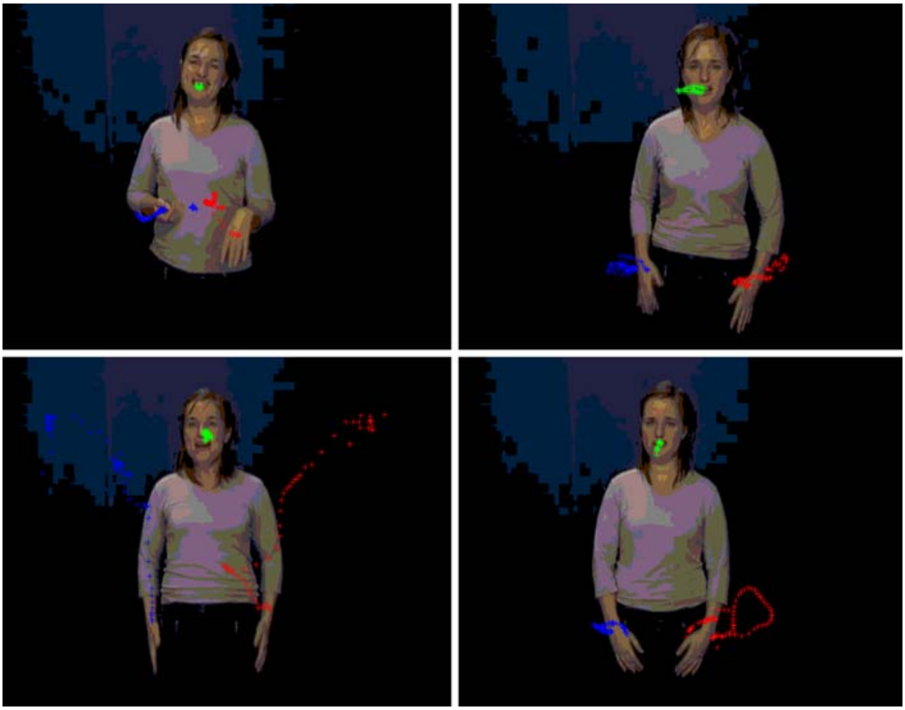

Fig. 10 Hand and head tracking results

misclassification is not propagated. Head and head occlusion is tackled with the following simplistic yet efficient method. Suppose occlusion occurs at frame n and ceases to exist at frame k. The position of the hand during the occlusion phase (frames $n-k$) is considered to be the position of the hand at frame $n-1$. After frame k the detection and tracking algorithm for the specific hand continues normally. Figures 10 and 11 demonstrates the hand and head trajectories produced by the proposed algorithm when applied to the GEMEP database (Baenziger et al. 2006) and the subjects of our own experimental corpus.

3.4 Gesture expressivity features calculation

Expressivity of behavior is an integral part of the communication process as it can provide information on the current emotional state, mood, and personality of a person (Wallbott and Scherer 1986). Many researchers have investigated human motion characteristics and encoded them into dual categories such as slow/fast, small/expansive, weak/energetic, small/large, and unpleasant/pleasant. To model expressivity, in our work, we use the six dimensions of behavior described in Hartmann et al. (2005a), as a more accomplished way to describe the expressivity, since it tackles all the parameters of expression of emotion. Five parameters

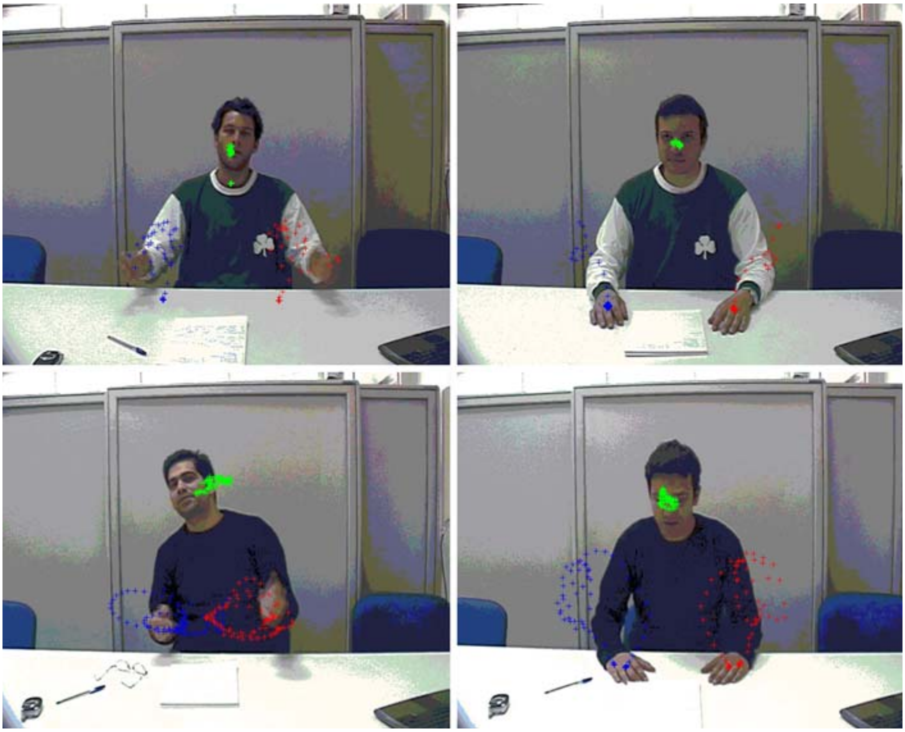

Fig. 11 Head and hand tracking results on scenario subjects

modeling behavior expressivity have been defined at the analysis level, as a subset of the above-mentioned six dimensions of behavior (see also next section):

- Overall activation
- Spatial extent
- Temporal
- Fluidity
- Power

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: $\mathrm{OA} = \sum_{i=0}^{n} |\vec{r}(i)| + \left|\vec{l}(i)\right|$. Spatial extent is modeled by expanding or condensing the entire space in front of the agent that is used for gesturing and is calculated as the maximum Euclidean distance of the position of the two hands: $\mathrm{SE} = \max\left(\left|d\left(\vec{r}(i) - \vec{l}(i)\right)\right|\right)$. The average spatial extent is also calculated for normalization reasons. The temporal expressivity parameter of the gesture signifies the duration of the movement while the speed expressivity parameter refers to the arm movement during the gesture's stroke phase (e.g., quick versus sustained actions). Gestures have three phases: preparation, stroke, and retraction. The real message is in the stroke, whilst the preparation and retraction elements consist of moving the arms to and from the rest position, to and from the start and end of the stroke. 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. Power actually is identical with the first derivative of the motion vectors calculated in the first steps.

4 Synthesis

Communicative capabilities of conversational agents could be significantly improved if they could also convey the expressive component of physical behavior. Starting from the results reported in Wallbott and Scherer (1986), we have defined and implemented (Hartmann et al. 2005a) a set of five parameters that affect the quality of the agent's behavior, that is the movement's spatial volume (SPC), speed (TMP), energy (PWR), fluidity (FLT), and repetitivity (REP). Thus, the same gestures or facial expressions are performed by the agent in a qualitatively different way depending on this set of parameters.

Table 1 shows the effect that each one of expressivity parameter has on the production of head movements, facial expressions 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 during the final animation. 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 in the forehead (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. On of the effect on the face is to speed up or slow down the rising/lowering of the eyebrows. The agent animation is generated by defining some key frames and computing the interpolation curves passing through these frames. 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/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. Thus, during its final animation the head will have a 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 reach the final position. The last parameter, Repetition (REP), exerts an influence on gestures and head movements. It increases the number of stroke of gestures to obtain repetition of the gestures themselves in

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

	Head	Facial expression	Gesture
SPC	Wider/narrower movement	Increased/decreased emotion arousal	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 movement	Increases/reduces continuity between consecutive gestures
PWR	Higher/shorter head overshooting	Higher/shorter movement acceleration	More/less stroke acceleration
REP	More/less number of nods and shakes	Not implemented yet	More/less number of repetitions of the same stroke

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.

Table 1 can be better understood with two intuitive examples. SPC affects the amplitude of the agent's head facial and body gestures: if a high value of SPC is selected and the agent has to perform a smile, the corners of her lips will widen and turn up the maximum. TMP affects the speed of head movements, facial expressions appearance and disappearance and gestures preparation and retraction. Then, for example, if a low value of TMP is selected and the agent has to nod, show a frown and perform a beat gesture, her head nod will be sluggish, her eyebrows will knit slowly, and she will move her arm slowly before the stroke of the beat gesture.

The synthesis module is able to reproduce a large set of facial expressions, the basic ones proposed by Ekman (1999) and many others obtained as a combination of them. Gestures are computed through the interpolation of a sequence of static positions obtained defining shoulder and arm rotation (arm position), hand shape (chosen in a set of predefined shapes) and palm orientation (Hartmann et al. 2002). So the synthesis module can successfully reproduce beat and iconic gestures whereas circular gestures cannot be performed.

From the point of view of implementation, our agent system produces animation data in MPEG4-compliant FAP/BAP format, which in turn drive a facial and skeletal body model in OpenGL. A set of parameters, called FAPs and Body Animation Parameters (BAPs), are used to animate the face and the body. By specifying values for FAPs and BAPs, we can specify facial expressions and body positions. The animation is specified by a sequence of keyframes. A keyframe is defined by a set of FAP or BAP values to be reached. Animation is obtained by interpolating between these keyframes. Interpolation is performed using TCB (Tencion, Continuity, Bias) splines (Kochanek and Bartels 1984).

The expressivity parameters are implemented by changing the TCB parameters of the interpolating splines and by scaling the values and changing the timing of the keyframes points. For example, the SPC parameter will influence the value of the keyframes by scaling them. The higher SPC will be, the wider the interpolating curves will be and so facial expressions will be more visible on the face and gestures wider. The FLD parameter will modulate the Continuity parameters of the splines, making them becoming smoother (high FLD) or jerkier (low FLD).

5 Implementation

In Sect. 1 we described the general framework of a system able to analyze a real scene and generate the animation of a virtual agent. Here we present a scenario that is a partial implementation of this framework. Currently our system (Fig. 12) is able to extract data from the video stream, process it and generate the animation of virtual agent. The final animation we aim to obtain consists on reproducing the gesture which is manually communicated to the agent and the facial expression that the system deduces from those performed by an actor.

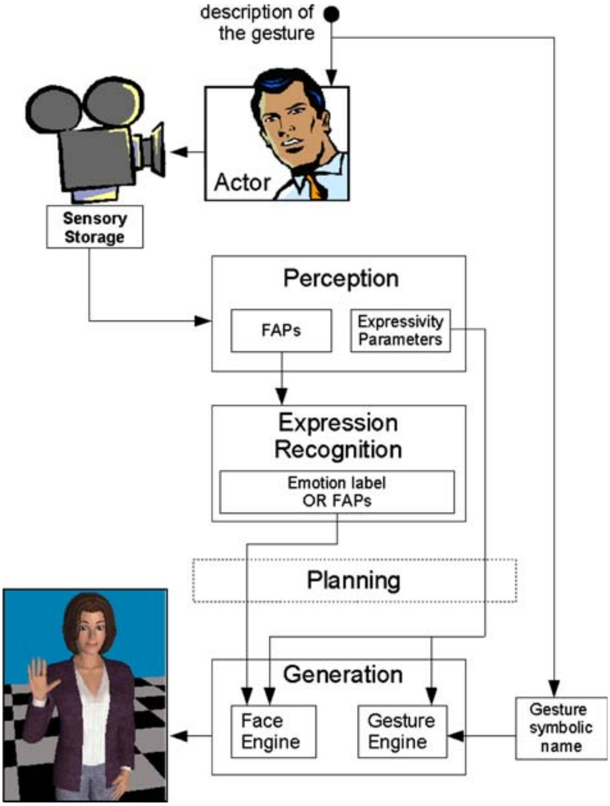

Fig. 12 Application scenario

The input is coming from an acted action performed by an actor. The action consists of a gesture accompanied by a facial expression. Both the type of the gesture and the type of the expression are explicitly requested from the actor and previously described to him in natural language in an acted portrayal scenario (Baenziger et al. 2006) (for example the actor is asked “to wave his right hand in front of the camera while showing a happy face”). Real life corpus was considered but was not selected because expressions sampled in real-life occur in specific contexts which often cannot be reproduced, their verbal content and the overall quality of the recordings cannot be controlled, one person is usually recorded in only one or very few different emotional states, and it is not always clear on which grounds and how these states should be labeled.

The *Perception* module analyzes 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 Expression Recognition 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 actual system is not able to extract the gesture shape from the data yet. In the near future, we would like also to implement a Planning module (represented in Fig. 12 with a dashed box) that could compute a modulation either of the expressivity parameters or the emotion. Finally the animation, consisting of variation of FAPs and BAPs values during time, is computed in the Generation module which contains the Face and the Gesture Engine. The Face Engine (which actually computes also the head movements of the agent) receives in input either the emotion label (or a list of FAPs) and a set of expressivity parameters. The way in which the facial expressions appear and head movements are performed is modulated by the expressivity parameters as explained in Sect. 4, Table 1. In the same way, the Gesture Engine receives in input a gesture label and a set of expressivity parameters. So the gestures produced by the Gesture Engine are influenced by the actual set of expressivity parameters, as explained in Sect. 4, Table 1.

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 goodness of movements.

The input image sequences of the presented system are videos captured at an acted session including 7 actors (Fig. 13), every one of them performing 7 gestures (Table 2). Each gesture was performed several times with the student-actor impersonating a different situation. Namely the gestures performed are: “*explain*,” “*oh my god*” (both hands over head), “*leave me alone*,” “*raise hand*” (draw attention), “*bored*” (one hand under chin), “*wave*,” “*clap*.” Table 2 indicates which emotion repetitions were performed for every specific gesture. For example gesture “*wave*” was performed 4 times one for the neutral emotion and three ((+, +), (−, +), (−, −)) for specific quadrants of the Whissel’s wheel. Some combinations were not included in the scenario since it did not make much sense in performing the “*bored*” gesture in a happy (+, +) mood.

Fig. 13 Subjects of the scenario

Table 2 Acted emotions

Gesture class	Quadrant of Whissel’s wheel
explain	(0,0), (+, +), (−, +), (−, −)
oh my god	(+, +), (−, +)
leave me alone	(−, +), (−, −)
raise hand	(0,0), (+, +), (−, −)
bored	(−, −)
wave	(0,0), (+, +), (−, +), (−, −)
clap	(0,0), (+, +), (−, +), (−, −)

Figure 14 shows the values of the expressivity features using the previously described algorithm on the gestures performed during the described experiment. Figure 15 proves the soundness of the overall module for calculating expressivity features by indicating the expected result. This would be that gestures belonging to the positive activation half-plane have higher scores on Overall Activation and Power in comparison to those belonging to the negative activation half-plane.

Figure 16 demonstrates three instances of behavior mimicry. The gesture being mimicked is “*raise hand*.” (a) instance is neutral where as (b) is happy (+/+ quadrant) and (c) is sad (−/− quadrant). Not all expressivity features can be observed from just an instance of the animation but Spatial Extent is very easily distinguishable for both the gestural and the facial aspect of the animation.

6 Evaluation scheme

A formative evaluation of the so far implemented system is judged of crucial importance in order to decipher how the synthesized behavior is perceived and thus identify ways to improve the current output produced. Feedback on expressivity makes sense both within and out of context (Hartmann et al. 2005b). At this point we are going to present a scheme on ways we propose in order to evaluate current work. Due to strict time limitations the results of this ongoing evaluation scheme will be made available in future publications.

There are various questions worth answering, through rating tests, regarding the perception of synthesized emotional expressions in order to continue an in depth analysis of the parameters just mentioned. In our case, a first question is the perception and classification of the synthesized output by human viewers. We plan to conduct a rating test where twenty postgraduate students will be asked to rate a sequence of videos—one at a time. They will be presented a questionnaire comprised of scale questions regarding the percent of each class they believe the

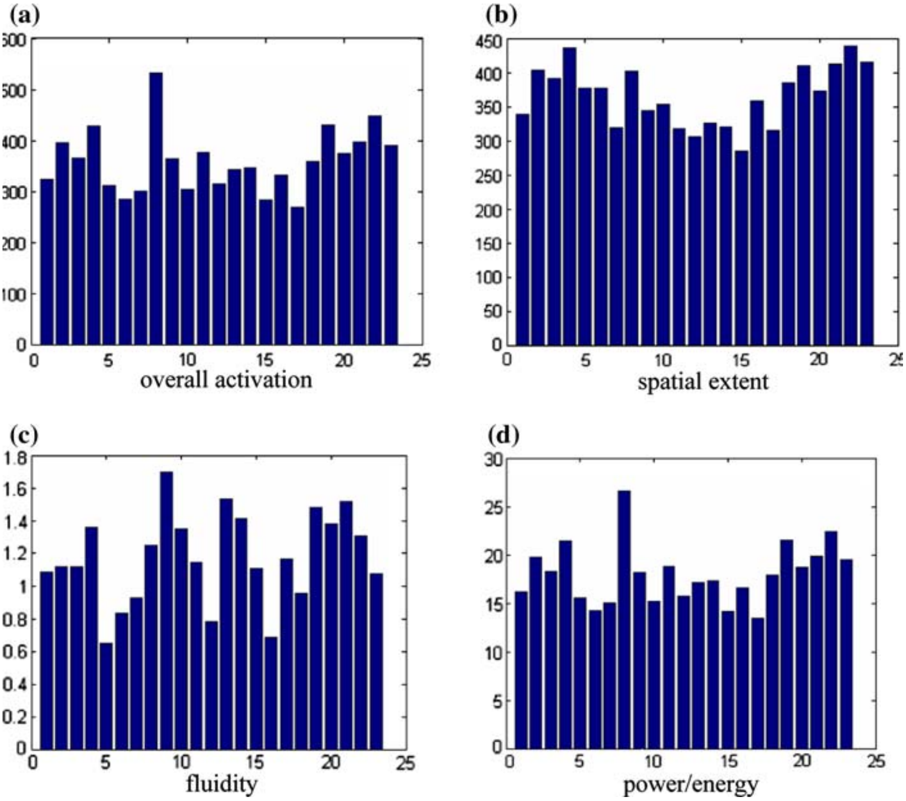

Fig. 14 Mean expressivity features values for every gesture class

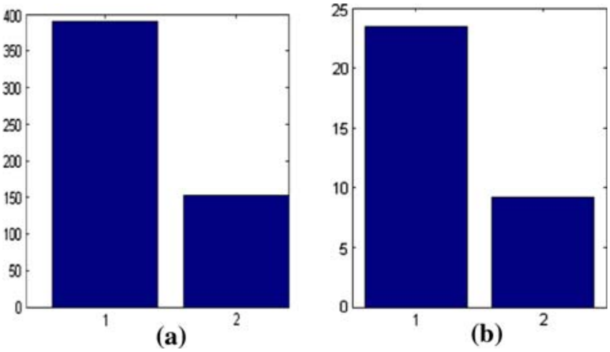

Fig. 15 Mean values of overall activation (**a**) and power (**b**) for positive and negative values of activation respectively

video snippet they just watched belongs to. The choice of classes is dictated by the natural language scenarios given to the actors of the original videos.

Participants will be asked how synthesized animations are perceived both in and out of the originating actor context by viewing the videos of the behavior mimicry

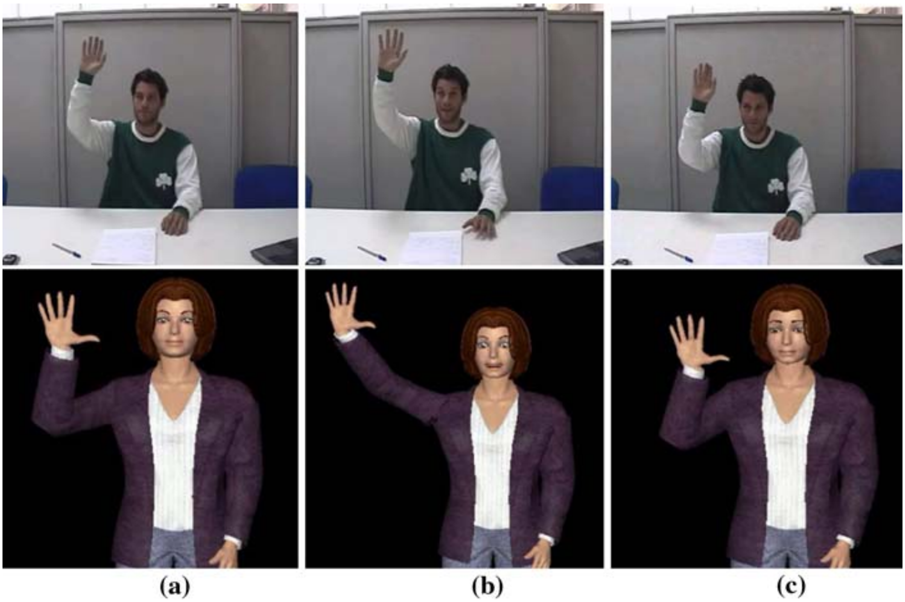

Fig. 16 Demonstration of behavior mimicry

scenarios. Thus, they will be divided into two groups, and will be randomly assigned to two conditions; the first group will be presented with the synthesized videos whereas the second group will be presented with both acted videos and their synthesized equivalents. The playback order of each set of videos will be chosen randomly so as to avoid ordering effects. In each group we plan to include a neutral condition of Greta simply blinking and looking straight ahead with a neutral expression as our conceptual baseline.

Questionnaires will also be used to collect participant feedback on which emotion they identified in each of the sequences they viewed. Their replies will be constrained with labels. Participants' confidence shall again be measured indirectly by asking them questions about the intensity of the perceived emotion.

Results from this first rating test can provide useful data on the perception and recognition of the synthesized expressions, as well as information on the effect of context (acted video sequences) in affective state perception of synthesized mimicked versions. Confidence measures will help draw conclusions on the role of the expressivity parameters and further refined manipulation of these parameters in conjunction with new rating tests can help decipher the role of these parameters in the perception of synthesized expressions.

7 Conclusions

We have presented our general framework consisting of a number of interconnected modules and one of its possible scenarios 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 comes from 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, while an extension which takes into account affective cues from the speech prosody channel is currently developed (Caridakis et al. 2007).

A very interesting extension to the framework is that of perceiving visual attention cues from the user (Rapantzikos and Avrithis 2005). As seen in the design of Fig. 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 behavior generation, such as the orienting of agent gaze. An attention system, applicable to both real and virtual environments, in a unified framework, is an interesting prospect. In addition to this, context information is an extremely important factor when trying to analyze the semantic underpinnings of human behavior; attention, user profiling and personalization and related adaptation processes are aspects that a planning component needs to take into account. Finally, we also aim to use the analysis-synthesis loop as a learning phase to refine the synthesis model of expressivity and of behavior.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274.

Baenziger, T., Pirker, H., & Scherer, K. (2006). Gemep - Geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions. In L. Devillers et al. (Eds.), *Proceedings of LREC'06 Workshop on corpora for research on emotion and affect* (pp. 15–19). Italy: Genoa.

Byun, M., & Badler, N. (2002). Facemote: Qualitative parametric modifiers for facial animations. In *Symposium on Computer Animation*, San Antonio, TX.

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., & Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. In *Proceedings of the 4th IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2007*, Athens, Greece.

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., & Karpouzis, K. (2006). Modeling naturalistic affective states via facial and vocal expressions recognition. In *International Conference on Multimodal Interfaces (ICMI'06)*, Banff, Alberta, Canada, November 2–4, 2006.

Chartrand, T. L., Maddux, W., & Lakin, J. (2005). Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. In R. Hassin, J. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 334–361). New York, NY: Oxford University Press.

Chi, D., Costa, M., Zhao, L., & Badler, N. (2000). The emote model for effort and shape. In *ACM SIGGRAPH '00*, pp. 173–182, New Orleans, LA.

Donato, G., Bartlett, M., Hager, J., Ekman, P., & Sejnowski, T. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(10), 974–989.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition & emotion* (pp. 301–320). New York: John Wiley.

Ekman, P., & Friesen, W. (1969). The repertoire of nonverbal behavioral categories – origins, usage, and coding. *Semiotica, 1*, 49–98.

Ekman, P. & Friesen, W. (1978). *The facial action coding system*. San Francisco, CA: Consulting Psychologists Press.

Hartmann, B., Mancini, M., & Pelachaud, C. (2002). Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. In *Computer Animation'02*, Geneva, Switzerland. IEEE Computer Society Press.

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

Hartmann, B., Mancini, M., Buisine, S., & Pelachaud, C. (2005b). Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *AAMAS'05.* Utretch.

Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T., Karpouzis, K., & Kollias, S. (2005). Emotion recognition through facial expression analysis based on a neurofuzzy network. *Special Issue on Emotion: Understanding & Recognition, Neural Networks, 18*(4), 423–435.

Juslin, P., & Scherer, K. (2005). Vocal expression of affect. In J. Harrigan, R. Rosenthal, & K. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research.* Oxford, UK: Oxford University Press.

Kochanek, D. H., & Bartels, R. H. (1984). Interpolating splines with local tension, continuity, and bias control. In H. Christiansen (Ed.), *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH' 84* (pp. 33–41). New York, NY: ACM. http://doi.acm.org/10.1145/800031.808575.

Kopp, S., Sowa, T., & Wachsmuth, I. (2003). Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures. In *Gesture Workshop*, pp. 436–447.

Lakin, J., Jefferis, V., Cheng, C., & Chartrand, T. (2003). The Chameleon effect as social Glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior, 27*(3), 145–162.

Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., & Pelachaud, C. (2005). 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, pp. 405–417.

Ong, S., & Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(6), 873–891.

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM, 42*(11), 74–81.

Pelachaud, C., & Bilvi, M. (2003). Computational model of believable conversational agents. In M.-P. Huget (Ed.), *Communication in multiagent system*s, Vol. 2650 of Lecture notes in *Computer Science* (pp. 300–317). Springer-Verlag.

Peters, C. (2005). Direction of attention perception for conversation initiation in virtual environments. In *International Working Conference on intelligent virtual agents*, Kos, Greece, pp. 215–228.

Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., & Kollias, S. (2002). Parameterized facial expression synthesis based on MPEG-4. *EURASIP Journal on Applied Signal Processing, 1*(Jan), 1021–1038. http://dx.doi.org/10.1155/S1110865702206149

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

Scherer, K., & Ekman, P. (1984). *Approaches to emotion*. Hillsdale: Lawrence Erlbaum Associates.

Tekalp, A., & Ostermann, J. (2000). Face and 2-d mesh animation in mpeg-4. *Signal Processing: Image Communication, 15*, 387–421.

van Swol, L. (2003) The effects of nonverbal mirroring on perceived persuasiveness, agreement with an imitator, and reciprocity in a group discussion. *Communication Research, 30*(4), 461–480.

Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology, 51*(4), 690–699.

Wexelblat, A. (1995). An approach to natural gesture in virtual environments. *ACM Transactions on Computer-Human Interaction, 2*, 179–200.

Whissel, C. M. (1989). The dictionary of affect in language. In R. Plutchnik & H. Kellerman (Eds.), *Emotion: Theory, research and experience: Vol. 4, The measurement of emotions*. New York: Academic Press.

Williams G. W. (1976). Comparing the joint agreement of several raters with another rater. *Biometrics, 32*, 619–627.

Wu, Y., & Huang, T. (2001). Hand modeling, analysis, and recognition for vision-based human computer interaction. *IEEE Signal Processing Magazine, 18*, 51–60.

Wu, Y., & Huang, T. S. (1999). Vision-based gesture recognition: A review. In *The 3rd gesture workshop*, Gif-sur-Yvette, France, pp. 103–115.