COMMUNICATION OVER THE INTERNET USING A 3D AGENT WITH REAL-TIME FACIAL EXPRESSION ANALYSIS, SYNTHESIS AND TEXT TO SPEECH CAPABILITIES

Naiwala P. Chandrasiri[†], István Barakonyi[†], Takeshi Naemura[‡], Mitsuru Ishizuka[†] and Hiroshi Harashima[†]

[†]University of Tokyo, School of Information Science & Technology, Tokyo, Japan.

Email: cds@hc.t.u-tokyo.ac.jp

[‡]Stanford University, Department of Computer Science, Stanford, CA, USA

***Abstract*— We present a system for Internet communication that enhances traditional text-based chatting with real-time analysis and synthesis of the chat parties' facial expressions. It is composed of three main modules: a real-time facial expression analysis component, a 3D agent with facial expression synthesis and text-to-speech capabilities – a talking head, and a communication module. So far we have realized a prototype to find out attractive ways of communication on the Internet and are currently experimenting on how to utilize this new type of modalities in chat communication.**

Keywords— facial expression, MPEG-4 FAP, 3D agent, chat

I. INTRODUCTION

THE evolution of the Internet has added new dimensions to communication among people who are geographically separated. Real-time chat applications such as ICQ and MSN messenger are two popular examples. They allow people to exchange text- or voice-based messages. Recently developed systems like FreeWalk [1] support casual meetings of small groups in a virtual space with real-time video and audio of the participants. In informal communication on the Internet, popularity of chat applications where one can appear by a nickname suggests that people can communicate more freely when they hide their identity. Also in some cases they prefer not to appear on the computer screen at all or only to show an avatar instead of them. On the other hand, communication without nonverbal information such as facial expressions, may be monotonous. Aiming at fulfilling both expectations, our system has the capabilities to animate agents based on real-time facial expression analysis techniques. This allows the user to be represented by a selected agent hiding his/her identity or as a predefined personalized agent. As the visual data exchange is based on MPEG-4 high-level expression Facial Animation Parameter (FAP 2) instead of real-time video, this method needs only a very low bandwidth.

This paper is mainly a combination of two previous works [2][3]. First work is responsible for calculating MPEG-4 high-level FAP 2 from facial images in real time and, the other one is responsible for synthesizing a 3D agent with facial expressions and text-to-speech capabilities. Interested reader may refer to them for further details.

II. RELATED WORK

There are already researches experimenting with enhanced chat communication utilizing talking heads [4][5][6], however, most of them don't consider facial expressions as an input channel. The background of our research includes facial animation, facial expression recognition and animated agents. Lifelike, animated agents with realistic behavior have gained much interest in the recent years. Their function and application domain is versatile; so far they have been utilized for instance as virtual actors [7], personal interactive tutors [8] and presentation agents [9]. Our agents are three-dimensional facial agents with synthetic speech, face and behavior; these agents represent the chat parties. From the several existing facial animation approaches [10] we chose the pseudomuscle-based concept with a face represented by a polygonal mesh. The pseudomuscle functions are conforming to action units of the Facial Action Coding System described in [11]. Besides facial expressions, head and eye motion is used to construct complex behavioral patterns. Moreover, appropriately changing facial colors help emphasize the agent's emotional state. In our 3D agent we didn't aim to achieve photorealism but rather easy modification and animation, in addition to reasonable speed and quality at today's average PC platform.

Using the MPEG-4 FAP, one can animate different 3D face models. Moreover, MPEG-4 high-level FAP 2 allows synthesizing various facial expressions with varying intensities (up to 2 primary categories' mixed expressions) [12]. However, the inverse problem of extracting MPEG-4 low- and high-level FAPs from real images is much more problematic due to the fact that the face is a highly deformable object. Furthermore, when it comes to recognition up to facial expression intensities in real time, subtle failures in facial area segmentation and various noise sources such as variation in lighting condition cause crucial problems.

For facial image synthesis applications, many approaches attempt to extract local spatial patterns such as action units (AU) and their combinations [13][14]. Real facial motion, however, is never completely localized. Detecting a unique set of action units for a specific facial expression is not guaranteed [14]. One promising approach for recognizing up to facial expression intensities is to consider the whole facial image as a single pattern. Kimura et al. has reported a method to construct emotion space using 2D elastic net model and K-L expansion for real images [15]. Their model is user independent, however, gives some undesirable results for unknown persons. Later, Ohba et al. also proposed facial expression space (FES) employing principal component analysis, considering the whole facial image as a single pattern [16]. FES is person dependant and eigen vectors are calculated for a specific person in the training mode. Then, one's input facial expression image is mapped into this eigen space. Their work is the most similar to our facial expression recognition method[2] in the sense that both are person dependant and consider the whole facial image as a single pattern. However, our method (Personal Facial Expression Space: PFES) is simpler which has the capability to construct it easily for any person or resetting it in a different environment such as different lighting conditions for the same person.

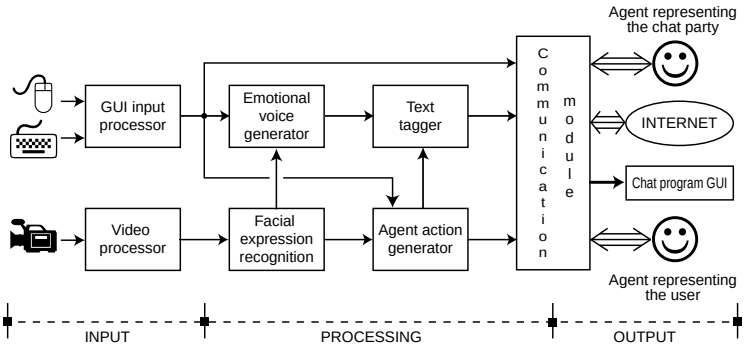


Fig. 1. System overview

III. System components

Our system (see Figure 1) takes various inputs from the chat party: the current face of the user by a head-mounted video camera and traditional GUI input like keyboard and mouse messages. Then several modules process the flow of input data. The facial expression recognition module calculates the MPEG-4 high-level FAP 2 of the current facial expression of the user. After the recognition, special tags for the text-to-speech synthesis engine are inserted into the chat message by the emotional voice generator module to artificially generate intonation carrying emotional content [9]. This adds another modality to the simple text information besides facial gestures and helps emphasize the chat message's emotional content. The recognized MPEG-4 high-level FAP 2 are also processed by the agent action generator, which decides on the appropriate animation command to be sent over to the agent representing the user. This module chooses the type and the intensity of the animation. These agent actions are immediately replayed by the local agent representation of the user providing some feedback how he/she behaves during the chat conversation. The agent animation commands and the tagged chat messages are then transmitted to the chat party over the Internet by the communication module. At the same time this module - as the channels are full duplex - processes data coming from other chat parties and passes it to the agents representing the chat parties to deliver their messages communicated by text, facial gestures and synthesized voice with emotional content.

A. Real-time facial expression recognition

In the process of estimating up to intensities of a primary facial expression image, we strongly believe that person specific modelling is unavoidable, as there is a considerable temporal and spatial variation of a facial expression and its apexes across individuals. One of the simplest ways of making person specific modelling is using apexes of primary facial expression images of that specific person as references. When an unknown input facial image from a primary facial image sequence appears, the problem is reduced to constructing a model with generalization capability using known reference images and, map the unknown input on it. For person specific modelling we acquire reference images which are the apexes of primary facial expressions and neutral face. We proposed a Personal Facial Expression Space (PFES) to recognize person specific, primary facial expression image sequences by both temporal and spatial characteristics taken into consideration [2]. To present a facial image, we define a global feature called expression vector, which is good enough to deal with real-time processing and noise factors. The PFES is constructed by using multidimensional scaling. Facial expression intensities of an input facial image which contains a mix of two primary facial categories can be easily calculated by its location on the PFES and it allows the calculation of the MPEG-4 high-level FAP 2. See Figure 2 (middle) for an example.

In designing a real-time system, two of the important issues which should be addressed are detection of a face and segmentation of facial area in the video input. The price we pay for the realization of a reliable real-time facial expression recognition system, however, is the need for a head mounted camera at this stage. We also designed a very small head mounted equipment (see Figure 2 (left)) by attaching a CMOS camera to a head phone. Currently, three facial expression categories – joy, surprise and anger – are used as they are relatively easy to be registered by the user. Figure 3 shows a screen shot of real-time mapping of a facial expression image onto PFES. Figure 3 (top) images are reference apex facial expressions which are registered in an interactive manner. We have prepared 2 masks which are based on an average face contour. By changing the camera settings and the mask type, the background image can be eliminated easily. Current input image (Figure 3 (right)) is mapped onto PFES (Figure 3 (bottom)). In MPEG-4 facial animation framework, two primary facial expressions can be animated simultaneously with amplitudes defined for each expression. We utilize 2D PFES considering the compatibility with MPEG-4 and, it's dynamically changed based on the expression vectors of the input image and reference images.

B. 3D facial agent

To synthesize the recognized facial expressions we use a 3D facial agent (see Figure 2 (right)) with customizable behavior and appearance [3]. The agent can both react upon explicit gesture commands (e.g. "make a nod" or "show a happy face"), and generate natural movements to provide the dynamic, realistic feature of the human face. These automatically generated natural movements include eye blinks, watering the mouth, natural rocking of the head, gazing around when idle or turning towards the listener when speaking and turning away when finishing the utterance. We referred to [17] and [18] for background work on this topic and we are extending them with our own studies. Lip sync to synthesized speech is done automatically as well. The agent features a

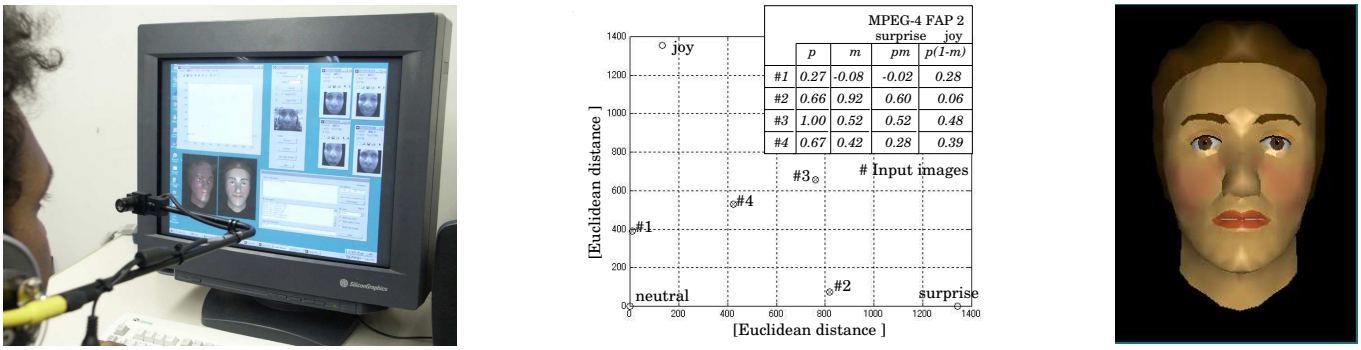


Fig. 2. (left) A user is on a chat session (middle) Real-time MPEG-4 high-level FAP 2 calculation (right) A screenshot of an agent.

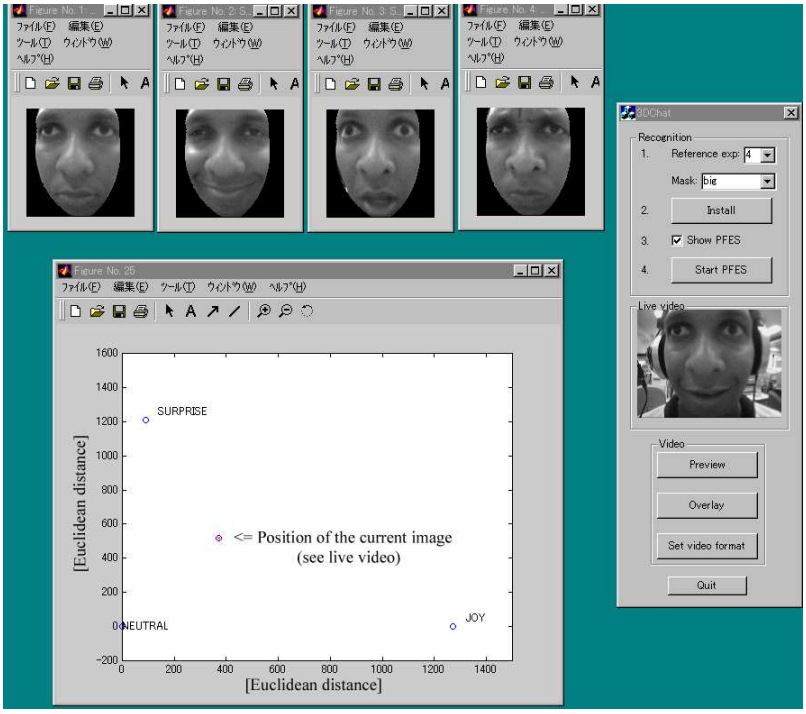


Fig. 3. Real time facial expression recognition module: (right) set up and live video window (top) registered reference images: neutral, joy, surprise and anger (bottom) current facial image position on PFES.

large set of animation commands, besides the 6 basic emotions [19]. There are ”ashamed”, ”suspicious”, ”disagreeing” etc. The user can arbitrarily extend this set by creating new animation actions, which can be downloaded by the chat partners and used in the communication as a new gesture if desired. However, these features have not been supported in the current system. These behavioral actions are composed of animation channels like facial expression, head and eye movement and facial color changes (e.g. blushing or pallor when changing the color of the cheek). The agent can also interpret MPEG-4 high-level FAP 2 as input command to generate appropriate facial expressions. The mesh representing the face is deformable, the user can choose from predefined faces during the chat session any time to change his/her appearance or can create arbitrary faces (including his/her own one) using a facial editor being developed. The chat parties can then download the new face and use it as the representation of the user when chatting. The agent is using a text-to-speech (TTS) engine to synthesize speech. By choosing from various TTS voices different agents sound in a different way representing sex, age and personality. Moreover, inserting special tags into the text to be uttered, the agent can communicate emotional content with the varying intonation [20]. Currently, speech synthesis are realized with Microsoft Speech API 5.0, thus all text-to-speech engines conforming to this interface can be used with the agent.

C. Communication module

The design of the Internet communication engine follows the client-server model, allowing for several chat parties, however, the current system is tested with two parties only, where depending on the situation both parties may become either server or client. Between the server and the client there are two communication channels maintained: text, and command channels. On the text channel messages typed by the chat party and tagged with agent animation information are transmitted, while the command channel is used for commands like synchronizing the two systems in case of possible command message losses, setting the chat parties' facial appearance or choosing between the chat modes.

A screen shot of a chat session is shown in Figure 4. While chatting, the user can select his/her face from set of predefined

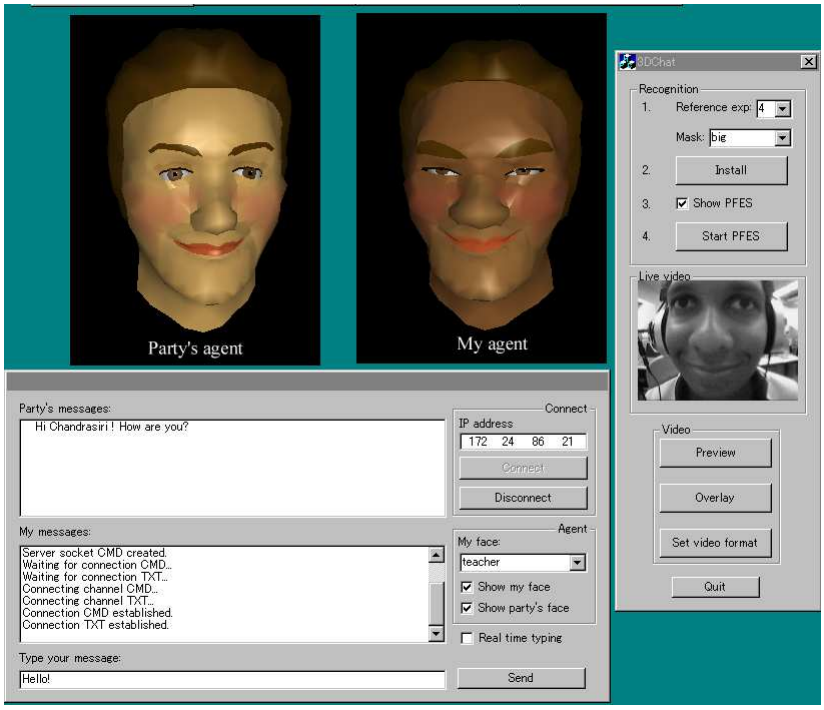


Fig. 4. A chat session: (top) agents (bottom) chat window (right) setup and live video window.

agents. Also there are options like changing real time typing mode which is discussed in the next section and showing/hiding agents.

In Figure 5, a 3D Agent (right) is synthesized by calculating MPEG-4 high level FAP 2 based on input facial image (left) in real time. This figure shows only three primary facial expressions and neutral, although our system is capable of synthesizing facial expressions with varying intensities.

Screen shots of an agent while speaking sentences with different facial expressions (and emotional content in the voice) are shown in Figure 6 (a)-(c). In our current system, while the remote party is typing a sentence, his/her facial expressions are analyzed and, if he/she presses a special key, his/her facial expression and its intensity at that time are inserted into the text (one emotion per one message segment).

IV. Discussion and future work

While we were testing our system, we tried several approaches to find out how to utilize the new modality of facial expression recognition when chatting. We created two modules. The first one exploits the real-time facial expression recognition and the other uses the full capabilities of the agent, e.g. speech synthesis and various facial gestures. However, there is a tradeoff between the two prototypes. For example, due to some restrictions in TTS, the agent needs a full sentence to speak out in a smooth way. If we utilize the agent's full capabilities, it can speak a sentence typed by a remote partner accompanied by a facial gestures and emotional content in the voice. However, you have to wait until the chat party types a full sentence, which may interrupt real-time communication. The second module allows real-time interaction in terms of typing and the behavior of the agent. This is realized by two channels which are dedicated to text and MPEG-4 high-level FAP 2 respectively. The information on the channels are synchronized so that you can see your remote partner typing a joke while smiling. At the same time you can react to it with a smile and some comments. Currently we are experimenting the existing prototypes as well as trying to find a hybrid solution where we can take advantage of the most features possible. We are also planning to conduct a psychological evaluation by asking users to try our system and tell us their impressions, likes and dislikes and desired future features. Further research will be conducted in order to add voice input and voice-to-viseme generation to supply the user with multiple inputs in the communication.

ACKNOWLEDGEMENTS

We would like to give thanks to Sylvain Descamps for providing us with constructive comments and criticism.

References

[1] H.Nakanishi, C. Yoshida, T. Nishimura, and T. Ishida, "Freewalk: 3D virtual space for casual meetings," *IEEE Multimedia*, vol. 6, no. 2, pp. 20–28, 1999.

[2] N. P. Chandrasiri, T. Naemura, and H.Harashima, "Real time facial expression recognition system with applications to facial animation in MPEG-4," *IEICE Transactions INF. & SYST.*, vol. E84-D, no. 8, pp. 1007–1017, 2001.

[3] I. Barakonyi and M. Ishizuka, "A 3D agent with synthetic face and semiautonomous behavior for multimodal presentations," *Proc. Multimedia Technology and Applications Conference (MTAC2001, IEEE Computer Soc.), Irvine, CA*, pp. 21–25, 2001.

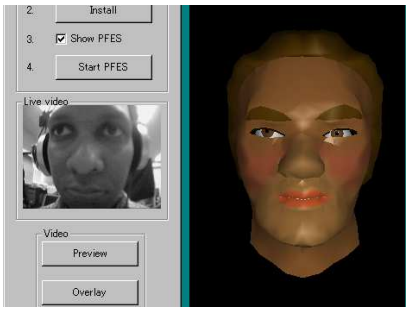
[4] L. C. De Silva, "Putting a face on the chat line," *Innovation*, vol. 2, no. 1, pp. 54–55, 2001.

[5] P. Hong, Z. Wen, and T. S. Huang, "IFACE: A 3D synthetic talking face," *International Journal of Image and Graphics*, vol. 1, no. 1, pp. 1–8, 2001.

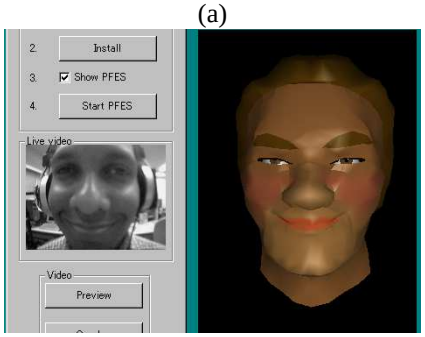
[6] S. Kshirsagar, C. Joslin, W. Lee, and N. Magnenat-Thalmann, "Personalized face and speech communication over the internet," *IEEE Virtual Reality*, pp. 37–44, 2001.

[7] K. Perlin and A. Goldberg, "Improv: A system for scripting interactive actors in virtual worlds," *Computer Graphics*, vol. 30, no. Annual Conference Series, pp. 205–216, 1996.

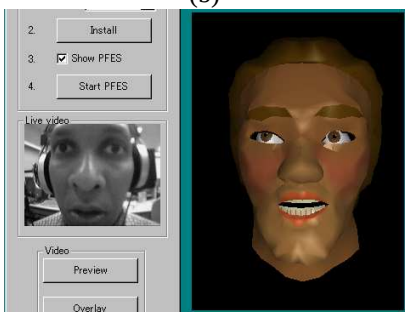
[8] J. Rickel and W. L. Johnson, "STEVE: A pedagogical agent for virtual reality," *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98), ACM Press, New York*, , no. 9–13, pp. 332–333, 1998.



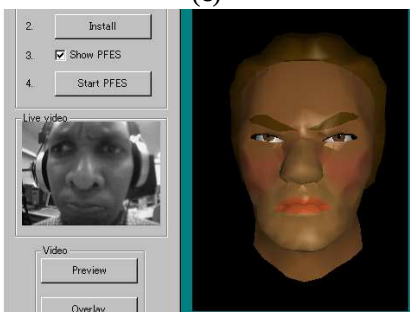
(a)



(b)

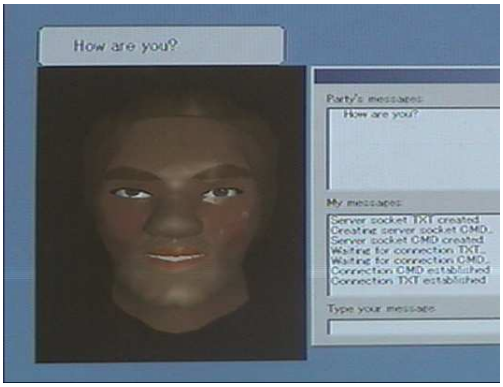


(c)

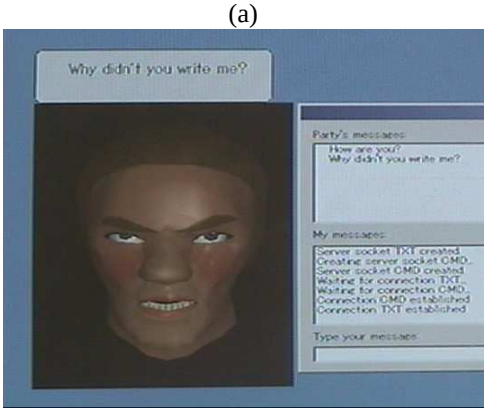


(d)

Fig. 5. (left) Input image (right) synthesized agent in real time. (a) neutral, (b) joy, (c) surprise and (d) anger.



(a)



(b)

(c)

Fig. 6. Agent is speaking sentences, typed by a remote party with different emotions: (a) ”How are you?”[joy], (b)”Why didn’t you write me? [angry]”, (c) ”Oh my God!”[surprise].

[9] S. Descamps, I. Barakonyi, and M. Ishizuka, “Making the web emotional: Authoring multimodal presentations using a synthetic 3D agent,” *OZCHI 2001 Conference Proceedings, Perth, Australia*, pp. 25–30, 2001.

[10] F. Parke and K. Waters, *Computer Facial Animation*, A K Peters, 1996.

[11] P. Ekman and W. Friesen, *Facial Action Coding System*, Consulting Psychologists Press, Inc., 1978.

[12] M. Tekalp and J. Osterman, “Face and 2-d mesh animation in mpeg-4,” *Signal Processing: Image Communication, Special Issue on MPEG-4*, vol. 15, pp. 387–421, 2000.

[13] C. S. Choi, K. Aizawa, H. Harashima, and T. Takebe, “Analysis and synthesis of facial image sequences in model-based image coding,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 4, pp. 257–275, 1994.

[14] I. A. Essa and A. P. Pentland, “Coding analysis, interpretation, and recognition of facial expressions,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757–763, 1997.

[15] M. Yachida S. Kimura, “Facial expression recognition and its degree estimation,” *Computer Vision and Pattern Recognition*, pp. 295–300, 1997.

[16] K. Ohba, G. Clary, T. Tsukada, T. Kotoku, and K. Tanie, “Facial expression communication with FES,” *International conference on Pattern Recognition*, pp. 1378–1378, 1998.

[17] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, “Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents,” *Computer Graphics*, vol. 28, no. Annual Conference Series, pp. 413–420, 1994.

[18] C. Pelachaud, N. Badler, and M. Steedman, “Generating facial expressions for speech,” *Cognitive Science*, vol. 20, no. 1, pp. 1–46, 1996.

[19] P. Ekman, “An argument for basic emotions,” *Cognition and Emotion*, vol. 6, pp. 169–200, 3–4 1992.

[20] J. Cahn, “The generation of affect in synthesized speech,” *Journal of the American Voice I/O Society, MIT Press*, vol. 8, pp. 1–19, 1990.