Face and Body Gesture Analysis for Multimodal HCI

Hatice Gunes, Massimo Piccardi, Tony Jan

Computer Vision Group, Faculty of Information Technology
University of Technology, Sydney (UTS), Australia
{haticeg, massimo, jant}@it.uts.edu.au

Abstract Humans use their faces, hands and body as an integral part of their communication with others. For the computer to interact intelligently with human users, computers should be able to recognize emotions, by analyzing the human's affective state, physiology and behavior. Multimodal interfaces allow humans to interact with machines through multiple modalities such as speech, facial expression, gesture, and gaze. In this paper, we present an overview of research conducted on face and body gesture analysis and recognition. In order to make human-computer interfaces truly natural, we need to develop technology that tracks human movement, body behavior and facial expression, and interprets these movements in an affective way. Accordingly, in this paper we present a vision-based framework that combines face and body gesture for multimodal HCI.

1 Introduction

In many HCI applications, for the computer to interact intelligently with human users, computers should be able to recognize emotions, by analyzing the human's affective state, physiology and behavior [23].

Non-verbal behavior plays an important role in human communications. According to Mehrabian [4], 93% of our communication is nonverbal and humans display their emotions most expressively through facial expressions and body gestures. Considering the effect of the message as a whole, spoken words of a message contribute only for 7%, the vocal part contributes 38%, while facial expression of the speaker contributes 55% to the effect of the spoken message [4]. Hence, understanding human emotions through nonverbal means is one of the necessary skills for computers to interact intelligently with their human counterparts. Furthermore, recent advances in image analysis and machine learning open up the possibility of automatic recognition of face and body gestures for affective human-machine communication [2,7].

This paper analyzes various existing systems and technologies used for automatic face and body gesture recognition and discusses the possibility of a multi-modal system that combines face and body signals to analyze the human emotion and behavior. The rationale for this attempt of combining face and body gesture for a better understanding of human non-verbal behavior is the recent interest and advances in multi-modal interfaces [24]. Pantic and Rothkrantz in [1] clearly state the importance of a multimodal affect analyzer for research in emotion recognition. The modalities considered are visual, auditory and tactile, where visual mainly stands for facial actions analysis. The interpretation of other visual cues such as body language (natural/spontaneous gestures) is not explicitly addressed in [1]. However, we think that this is an important component of affective communication and this will be a major goal in this paper. Moreover, an automated system that senses, processes, and interprets the combined modes of facial expression and body gesture has great potential in various research and application areas [1-3] including human-computer interaction and pervasive perceptual man-machine interfaces [23,24].

The paper is organized as follows. Section 2 and 3 cover previous work done on automatic facial expression and gesture analysis, respectively. Section 4 presents the possible efforts toward multimodal analyzers of human affective state. Section 5 presents our approach on combining face and body gesture for multimodal HCI and finally, Section 6 gives the conclusion.

2 Facial Expression

Facial expression measurement provides an indicator of emotion activity and is presently used in a variety of areas from behavioral research to HCI. Research in psychology has indicated that at least six emotions are universally associated with distinct facial expressions: happiness, sadness, surprise, fear, anger, and disgust [5,6]. Several other emotions, and many combinations of emotions have been studied but remain unconfirmed as universally distinguishable.

Facial Expression Recognition

Within the past decade, analysis of human facial expression has attracted great interest in the machine vision and artificial intelligence communities. Systems that automatically analyze the facial expressions can generally be classified into two categories: (1) systems that recognize prototypic facial expressions (happy, sad etc.); (2) systems that recognize facial actions (frown, eyebrow raise etc.).

***(1) Systems that Recognize Prototypic Facial Expressions**:* There has been a significant amount of research on creating systems that recognize a small set of prototypic emotional expressions from static images or image sequences. This focus follows from the work of Ekman [5]. Bassili suggested that motion in the image of a face would allow emotions to be identified even with minimal information about the spatial arrangement of features [6]. Thus, facial expression recognition from image sequences can be based on categorizing prototypic facial expressions by tracking facial features and measuring the amount of facial movement; various approaches have been explored [9, 10].

***(2) Systems that Recognize Facial Actions**:* Although prototypic expressions are natural, they occur infrequently in everyday life and provide an incomplete description of facial expressions. To capture the subtlety of human emotion and paralinguistic communication, automated recognition of fine-grained changes in facial expression is needed [11]. Ekman and Friesen developed the Facial Action Coding System (FACS) for describing facial expressions by action units (AUs) [5]. The system is based on the enumeration of all "action units" of a face that cause facial movements. Out of the 44 AUs defined, 30 are anatomically related to the contractions of specific facial muscles: 12 for the upper face, and 18 for the lower face. AUs can be classified either individually or in combination. Vision-based systems attempting to recognize action units (AUs) are motivated by FACS [5]. Some of them have used optical flow across the entire face or facial feature measurements

[9, 14]. Tian and Kanade used facial features to recognize 16 AUs and their combination by describing the shape of facial features by multistate templates [7]. Donato *et al*. compared optical flow, principal component analysis, independent component analysis, local feature analysis, and Gabor wavelet representation to recognize 6 single upper face AUs and 6 single lower face AUs [8]. Pantic and Rothkrantz developed a facial gesture recognition system from both frontal and profile image sequences [12]. Kapoor and Picard used pupil detection from infrared sensitive cameras for recognizing upper AUs [13].

3 Gesture

Gesture is "the use of motions of the limbs or body as a mean of expression; communicate an intention or feeling" [14]. Gestures include body movements (e.g., palm-down, shoulder-shrug) and postures (e.g., angular distance), and most often occur in conjunction with speech. It has been shown that when speech is ambiguous or in a noisy environment, listeners rely on the gestural cues [3]. Thus, gestures serve an important communicative function in face-to-face communication [3, 16]. The majority of hand gestures produced by speakers are connected to speech. Kendon has situated these hand gestures along a "gesture continuum" [14], defining five different kinds of gestures: (1)*Gesticulation*– spontaneous movements of the hands and arms that accompany speech; (2)*Language-like gestures*– gesticulation that is integrated into the vocal expression, replacing a particular spoken word; (3)*Pantomimes*– gestures that represent objects or actions, with or without accompanying speech; (4)*Emblems*– familiar gestures such as thumbs up (often culturally specific); *(5)Sign languages*– The well defined linguistic systems, such as Australian Sign Language.

Gesture Recognition

According to Turk [17], most research to date in computer science, focuses on emblems and sign languages in Kendon's continuum, where gestures tend to be less ambiguous, more learned and more culture-specific. However, the concept of gesture is still not well defined, and depends on the context of the particular interaction (static/dynamic). Currently, most computer vision systems for recognizing gestures have similar components [17]. Human position and movement is sensed using cameras and computer vision techniques. In the preprocessing stage images are normalized, enhanced, or transformed. Gestures are modeled by using various representation techniques [18, 19]. Features are extracted and gestures analyzed with motion analysis, segmentation, contour representation etc. [17-19]. In the final stage, gesture recognition and classification is performed.

4 Multimodal HCI

Multimodal systems provide the possibility of combining different modalities (i.e. speech, facial expression, gesture, and gaze) that occur together to function in a more efficient and reliable way in various human-computer interaction applications [1], [24]. Studies showed that these interfaces support more effective human-computer interaction [24]. Currently, there are very few multi-modal systems introduced attempting to analyze combinations of communication means for *human affective state analysis* [20-22]. These are systems mostly combining auditory and visual information by processing facial expression and vocal cues for affective emotion recognition. According to a recent survey by Pantic and Rothkrantz in [1], the work presented by Picard *et al.* [23] is the only work combining different modalities for automatic analysis of *affective physiological signals*. See [1] and [24] for further review of the recent attempts at combining different modalities in HCI.

5 Proposed Framework

Face and body gestures are two of the several channels of nonverbal communication that occur together. Messages can be expressed through face and gesture in many ways. For example, an emotion such as sadness can be communicated through the facial expression and lowered head and shoulder position. Thus, various nonverbal channels can be combined for construction of computer systems that can *affectively* communicate with humans.

We propose a vision-based framework that uses computer vision and machine learning techniques to recognize face and body gesture for a multimodal HCI interface. To our best knowledge there has been no attempt to combine face and body gesture for nonverbal behavior recognition. For our multimodal analyzer we will use a human model including the face (eyes, eyebrows, nose, lips and chin) and the upper body (trunk, two arms and two hands) performing face and body actions (i.e. raising eyebrows, crossing arms etc.). Hence, multi-modality will be achieved by combining facial expression and body language. Proposed system framework is shown in Fig.1.

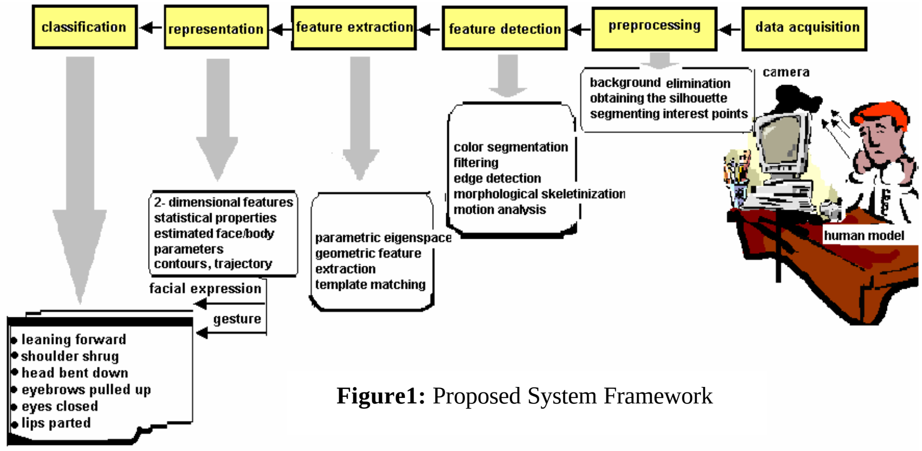

Figure1: Proposed System Framework

Due to being a new research area, there exist problems to be solved and issues to be considered in order to develop a robust multimodal analyzer of face and body gesture. In a multimodal system, the basic problem to be solved is to fuse information from different modalities [1]. Fusion could be (a) done early or late in the interpretation process; (b) some mode could be principal and other auxiliary. Another potential issue to consider in our work is that gesture analysis is even more context-dependent than facial action analysis. For this reason, we clearly want to treat face and body information separately. Another issue to consider is that detection of gesture actions could be technically more challenging than facial actions. There is a greater intrinsic visual complexity; facial features never occlude each other and they are not deformable, instead, limbs are subject to occlusions and deformations. However, the use of gesture actions could be an auxiliary mode to be used only when expressions from the remaining modes are classified as ambiguous.

6 Conclusion

This paper presented a vision-based framework that recognizes face and body gesture for multimodal HCI by firstly presenting various approaches and previous work in automatic facial expression/action analysis, gesture recognition and multimodal interfaces. "The advent of multimodal interfaces based on recognition of human speech, gaze, gesture, and other natural behavior represents only the beginning of a progression toward computational interfaces capable of human-like sensory perception" [24]. However, due to being a fairly new research area, there still exist problems to be solved and issues to be considered [1,24].

References

[1] M. Pantic and L.J.M. Rothkrantz, 'Towards an Affect-Sensitive Multimodal Human-Computer Interaction '. In: *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390, September 2003

[2] M. Pantic and L.J.M. Rothkrantz,, "Automatic analysis of facial expressions: the state of the art", *IEEE PAMI*, pp. 1424-1445, Vol: 22, Issue: 12, Dec 2000

[3] J. Cassell. A framework for gesture generation and interpretation. In R. Cipolla and A. Pentland, editors, *Computer vision in human-machine interaction*. Cambridge University Press, 2000.

[4] A. Mehrabian, "Communication without words," *Psychol. Today*, vol. 2, no. 4, pp. 53–56, 1968.

[5] P. Ekman and W. V. Friesen. *The Facial Action Coding System*. Consulting Psychologists Press, San Francisco, CA, 1978.

[6] J. N. Bassili. Facial motion in the perception of faces and of emotional expression , *Experimental Psychology*, 4:373–379, 1978.

[7] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE PAMI*, 23(2), February 2001.

[8] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE PAMI*, 21(10):974–989, October 1999.

[9] I. Essa and A. Pentland. Coding, analysis, interpretation and recognition of facial expressions, *IEEE PAMI*, 7:757–763, July 1997.

[10] K. Mase. Recognition of facial expressions for optical flow. IEICE Transactions, Special Issue on Computer Vision and its Applications, E 74(10), 1991.

[11] Joseph C. Hager and Paul Ekman, Essential Behavioral Science of the Face and Gesture that Computer Scientists Need to Know, 1995

[12] M. Pantic, I. Patras and L.J.M. Rothkrantz, 'Facial action recognition in face profile image sequences ', in *Proc. IEEE Int'l Conf. on Multimedia and Expo*, vol. 1, pp. 37-40, Lausanne, Switzerland, August 2002

[13] A. Kapoor and R. W. Picard. Real-time, fully automatic upper facial feature tracking, In Proc. of FG, May 2002.

[14] A. Kendon. How gestures can become like words. In F. Poyatos, editor, *Cross-cultural perspectives in nonverbal communication*, New York, 1988. C.J. Hogrefe.

[16] D. McNeill, (1985). So you think gestures are nonverbal? *Psychological Review*, *92*, 350-371.

[17] M. Turk, "Gesture Recognition," in *Handbook of Virtual Environments : Design, Implementation, and Applications*, K. Stanney, Ed.: Lawrence Erlbaum Associates, Inc.(Draft version), 2001.

[18] H. Ohno and M. Yamamoto, Gesture Recognition using Character Recognition Techniques on Two-Dimensional Eigenspace, Proc. of ICCV 1999, pp. 151-156

[19] P. Peixoto, J. Gonçalves, H. Araújo, Real-Time Gesture Recognition System Based on Contour Signatures", *ICPR'2002*, Quebec City, Canada, August 11-15, 2002

[20] L. S. Chen and T. S. Huang, "Emotional expressions in audiovisual human computer interaction," in *Proc. ICME*, 2000, pp. 423–426.

[21] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. FG*, 2000, pp. 332–335.

[22] Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," in *Proc. ROMAN*, 2000, pp. 178–183.

[23] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE PAMI.*, vol. 23, pp. 1175–1191, Oct. 2001.

[24] J.L. Franagan,T.S. Huang, "Scanning the issue special issue on human-computer multimodal interface", *In Proceedings of the IEEE*, Vol: 91, no: 9, pp. 1267- 1271, Sept. 2003