Face for Interface

Maja Pantic
Imperial College London, UK
University of Twente, The Netherlands

INTRODUCTION: THE HUMAN FACE

The human face is involved in an impressive variety of different activities. It houses the majority of our sensory apparatus: eyes, ears, mouth, and nose, allowing the bearer to see, hear, taste, and smell. Apart from these biological functions, the human face provides a number of signals essential for interpersonal communication in our social life. The face houses the speech production apparatus and is used to identify other members of the species, to regulate the conversation by gazing or nodding, and to interpret what has been said by lip reading. It is our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression (Lewis & Haviland-Jones, 2000). Personality, attractiveness, age, and gender can also be seen from someone's face. Thus the face is a multisignal sender/receiver capable of tremendous flexibility and specificity. In general, the face conveys information via four kinds of signals listed in Table 1.

Automating the analysis of facial signals, especially rapid facial signals, would be highly beneficial for fields as diverse as security, behavioral science, medicine, communication, and education. In security contexts, facial expressions play a crucial role in establishing or detracting from credibility. In medicine, facial expressions are the direct means to identify when specific mental processes are occurring. In education, pupils' facial expressions inform the teacher of the need to adjust the instructional message.

As far as natural user interfaces between humans and computers (PCs/robots/machines) are concerned, facial expressions provide a way to communicate basic information about needs and demands to the machine. In fact, automatic analysis of rapid facial signals seem to have a natural place in various vision subsystems and vision-based interfaces (face-for-interface tools), including automated tools for gaze and focus of attention tracking, lip reading, bimodal speech processing, face/visual speech synthesis, face-based command issuing, and facial affect processing. Where the user is looking (i.e., gaze tracking) can be effectively used to free computer users from the classic keyboard and mouse. Also, certain facial signals (e.g., a wink) can be associated with certain commands (e.g., a mouse click) offering an alternative to traditional keyboard and mouse commands. The human capability to "hear" in noisy environments by means of lip reading is the basis for bimodal (audiovisual) speech processing that can lead to the realization of robust speech-driven interfaces. To make a believable "talking head" (avatar) representing a real person, tracking the person's facial signals and making the avatar mimic those using synthesized speech and facial expressions is compulsory. The human ability to read emotions from someone's facial expressions is the basis of facial affect processing that can lead to expanding user interfaces with emotional communication and, in turn, to obtaining a more flexible, adaptable, and natural affective interfaces between humans and machines. More specifically, the information about when the existing interaction/processing should be adapted, the importance of such an adaptation, and how the interaction/ reasoning should be adapted, involves information about how the user feels (e.g., confused, irritated, tired, interested). Examples of affect-sensitive user interfaces are still rare, unfortunately, and include the systems of Lisetti and Nasoz (2002), Maat and Pantic (2006), and Kapoor, Burleson, and Picard (2007). It is this wide range of principle driving applications that has lent a special impetus to the research problem of automatic facial expression analysis and produced a surge of interest in this research topic.

BACKGROUND: FACIAL ACTION CODING

Rapid facial signals are movements of the facial muscles that pull the skin, causing a temporary distortion of the shape of the facial features and of the appearance of folds, furrows, and bulges of skin. The common terminology for describing rapid facial signals refers either to culturally dependent linguistic terms

F

Table 1. Four types of facial signals

• *Static facial signals* represent relatively permanent features of the face, such as the bony structure, the soft tissue, and the overall proportions of the face. These signals are usually exploited for person identification.
• *Slow facial signals* represent changes in the appearance of the face that occur gradually over time, such as development of permanent wrinkles and changes in skin texture. These signals can be used for assessing the age of an individual.
• *Artificial signals* are exogenous features of the face such as glasses and cosmetics. These signals provide additional information that can be used for gender recognition.
• *Rapid facial signals* represent temporal changes in neuromuscular activity that may lead to visually detectable changes in facial appearance, including blushing and tears. These (atomic facial) signals underlie facial expressions.

indicating a specific change in the appearance of a particular facial feature (e.g., smile, smirk, frown) or to the linguistic universals describing the activity of specific facial muscles that caused the observed facial appearance changes.

There are several methods for linguistically universal recognition of facial changes based on the facial muscular activity (Scherer & Ekman, 1982). From those, the facial action coding system (FACS) proposed by Ekman and Friesen (1978) and Ekman, Friesen, and Hager (2002) is the best known and most commonly used system. It is a system designed for human observers to describe changes in the facial expression in terms of visually observable activations of facial muscles. The changes in the facial expression (rapid facial signals) are described with FACS in terms of 44 different action units (AUs), each of which is anatomically related to the contraction of either a specific facial muscle or a set of facial muscles. Examples of different AUs are given in Table 2. Along with the definition of various AUs, FACS also provides the rules for visual detection of AUs and their temporal segments (onset, apex, offset) in a face image. Using these rules, a FACS coder (that is, a human expert having a formal training in using FACS) decomposes a shown facial expression into the AUs that produce the expression.

Although FACS provides a good foundation for AU-coding of face images by human observers, achieving AU recognition by an automated system for facial expression analysis is by no means a trivial task. A problematic issue is that AUs can occur in more than 7,000 different complex combinations (Scherer & Ekman, 1982), causing bulges (e.g., by the tongue pushed under one of the lips) and various in- and out-of-image-plane movements of permanent facial features (e.g., jetted jaw) that are difficult to detect in 2D face images.

AUTOMATED FACIAL ACTION CODING

Most approaches to automatic facial expression analysis attempt automatic facial affect analysis by recognizing a small set of prototypic emotional facial expressions, that is, fear, sadness, disgust, anger, surprise, and happiness, (For exhaustive surveys of the past work on this research topic, readers are referred to the work of Pantic & Rothkrantz, 2003, and Zeng, Pantic, Roisman, & Huang, 2007.) This practice may follow from the work of Darwin and more recently Ekman (Lewis & Haviland-Jones, 2000), who suggested that basic emotions have corresponding prototypic expressions. In everyday life, however, such prototypic expressions occur relatively rarely; emotions are displayed more often by subtle changes in one or few discrete facial features, such as raising of the eyebrows in surprise. To detect such subtlety of human emotions and, in general, to make the information conveyed by facial expressions available for usage in the various applications mentioned above, including user interfaces, automatic recognition of rapid facial signals (AUs) is needed.

A number of approaches have been reported up to date for automatic recognition of AUs in images of faces. For exhaustive surveys of the related work, readers are referred to the work of Tian, Kanade, and Cohn (2005), Pantic (2006), and Pantic and Bartlett (2007). Some researchers described patterns of facial motion that correspond to a few specific AUs, but did

Related Content