

Face for Interface

Maja Pantic

Delft University of Technology, 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; it regulates conversation by gazing or nodding and interprets 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 also can be seen from someone's face. Thus, the face is a multi-signal 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 medi-

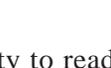
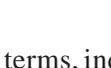
cine, facial expressions are the direct means to identify when specific mental processes are occurring. In education, pupils' facial expressions inform the teacher of the need to adjust the instructional message.

As far as natural interfaces between humans and computers (i.e., 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 seems to have a natural place in various vision subsystems, 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 are com-

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*.

Table 2. Examples of facial action units (AUs)

	AU1: Raised inner eyebrow		AU2: Raised outer eyebrow
	AU1 + AU2: Raised eyebrows		AU4: Lowered eyebrow Eyebrows drawn together
	AU5: Raised upper eyelid		AU6: Raised cheek Compressed eyelid
	AU7: Tightened eyelid		AU41: Drooped eyelid
	AU44: Squinted eyes		AU46: Wink
	AU9: Wrinkled nose		AU11: Deepened nasolabial furrow
	AU12: Lip corners pulled up		AU13: Lip corners pulled up sharply
	AU14: Dimpler - mouth corners pulled inwards		AU15: Lip corners depressed
	AU17: Chin raised		AU19: Tongue shown
	AU20: Mouth stretched horizontally		AU24: Lips pressed
	AU26: Jaw dropped		AU29: Jaw pushed forward
	AU30: Jaw sideways		AU36: Bulge produced by the tongue

pulsory. The human ability to read emotions from someone's facial expressions is the basis of facial affect processing that can lead to expanding interfaces with emotional communication and, in turn, obtain a more flexible, adaptable, and natural interaction between humans and machines.

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, indicating a specific change in the appearance of a particular facial feature (e.g., smile, smirk, frown, sneer), or for 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 et al. (1978, 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 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 (i.e., onset, apex, offset) in a face image. Using these rules, a FACS coder (i.e., a human expert having 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 a computer 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 to recognize a small set of prototypic emotional facial expressions (i.e., fear, sadness, disgust, anger, surprise, and happiness) (for an exhaustive survey of the past work on this research topic, the reader is referred to the work of Pantic & Rothkrantz [2003]). 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 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, automatic recognition of rapid facial signals (AUs) is needed.

Few approaches have been reported for automatic recognition of AUs in images of faces. Some researchers described patterns of facial motion that correspond to a few specific AUs, but did not report on actual recognition of these AUs. Examples of such works are the studies of Mase (1991) and Essa and Pentland (1997). Almost all other efforts in automating FACS coding addressed the problem of automatic AU recognition in face video using both machine vision techniques like optical flow analysis, Gabor wavelets,

temporal templates, particle filtering, and machine learning techniques such as neural networks, support vector machines, and hidden Markov models. To detect six individual AUs in face image sequences free of head motions, Bartlett et al. (1999) used a neural network. They achieved 91% accuracy by feeding the pertinent network with the results of a hybrid system combining holistic spatial analysis and optical flow with local feature analysis. To recognize eight individual AUs and four combinations of AUs with an average recognition rate of 95.5% for face image sequences free of head motions, Donato et al. (1999) used Gabor wavelet representation and independent component analysis. To recognize eight individual AUs and seven combinations of AUs with an average recognition rate of 85% for face image sequences free of head motions, Cohn et al. (1999) used facial feature point tracking and discriminant function analysis. Tian et al. (2001) used lip tracking, template matching, and neural networks to recognize 16 AUs occurring alone or in combination in nearly frontal-view face image sequences. They reported an 87.9% average recognition rate attained by their method. Braathen et al. (2002) reported on automatic recognition of three AUs using particle filtering for 3D tracking, Gabor wavelets, support vector machines, and hidden Markov models to analyze an input face image sequence having no restriction placed on the head pose. To recognize 15 AUs occurring alone or in combination in a nearly frontal-view face image sequence, Valstar et al. (2004) used temporal templates. Temporal templates are 2D images constructed from image sequences, which show where and when motion in the image sequence has occurred. The authors reported a 76.2% average recognition rate attained by their method.

In contrast to all these approaches to automatic AU detection, which deal only with frontal-view face images and cannot handle temporal dynamics of AUs, Pantic and Patras (2004) addressed the problem of automatic detection of AUs and their temporal segments (onset, apex, offset) from profile-view face image sequences. They used particle filtering to track 15 fiducial facial points in an input face-profile video and temporal rules to recognize temporal segments of 23 AUs occurring alone or in a combination in the input video sequence. They achieved an 88% average recognition rate by their method.

The only work reported to date that addresses automatic AU coding from static face images is the work of Pantic and Rothkrantz (2004). It concerns an automated system for AU recognition in static frontal- and/or profile-view color face images. The system utilizes a multi-detector approach for facial component localization and a rule-based approach for recognition of 32 individual AUs. A recognition rate of 86% is achieved by the method.

CRITICAL ISSUES

Facial expression is an important variable for a large number of basic science studies (in behavioral science, psychology, psychophysiology, psychiatry) and computer science studies (in natural human-machine interaction, ambient intelligence, affective computing). While motion records are necessary for studying temporal dynamics of facial behavior, static images are important for obtaining configurational information about facial expressions, which is essential, in turn, for inferring the related meaning (i.e., in terms of emotions) (Scherer & Ekman, 1982). As can be seen from the survey given above, while several efforts in automating FACS coding from face video have been made, only Pantic and Rothkrantz (2004) made an effort for the case of static face images.

In a frontal-view face image (portrait), facial gestures such as showing the tongue (AU 19) or pushing the jaw forwards (AU 29) represent out-of-image-plane, non-rigid facial movements that are difficult to detect. Such facial gestures are clearly observable in a profile view of the face. Hence, the usage of face-profile view promises a qualitative enhancement of AU detection performed by enabling detection of AUs that are difficult to encode in a frontal facial view. Furthermore, automatic analysis of expressions from face profile-view would facilitate deeper research on human emotion. Namely, it seems that negative emotions (where facial displays of AU2, AU4, AU9, and the like are often involved) are more easily perceivable from the left hemiface than from the right hemiface, and that, in general, the left hemiface is perceived to display more emotion than the right hemiface (Mendolia & Kleck, 1991). However, only Pantic and Patras (2004) made an effort to date to automate FACS

coding from video of profile faces. Finally, it seems that facial actions involved in spontaneous emotional expressions are more symmetrical, involving both the left and the right side of the face, than deliberate actions displayed on request. Based upon these observations, Mitra and Liu (2004) have shown that facial asymmetry has sufficient discriminating power to significantly improve the performance of an automated genuine emotion classifier. In summary, the usage of both frontal and profile facial views and moving toward 3D analysis of facial expressions promises, therefore, a qualitative increase in facial behavior analysis that can be achieved. Nevertheless, only Braathen et al. (2002) made an effort to date in automating FACS coding using a 3D face representation.

There is now a growing body of psychological research that argues that temporal dynamics of facial behavior (i.e., timing, duration, and intensity of facial activity) is a critical factor for the interpretation of observed behavior (Lewis & Haviland-Jones, 2000). For example, Schmidt and Cohn (2001) have shown that spontaneous smiles, in contrast to posed smiles, are fast in onset, can have multiple AU12 apexes (i.e., multiple rises of the mouth corners), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within one second. Hence, it is obvious that automated tools for the detection of AUs and their temporal dynamics would be highly beneficial. However, only Pantic and Patras (2004) reported so far on an effort to automate the detection of the temporal segments of AUs in face image sequences.

None of the existing systems for facial action coding in images of faces is capable of detecting all 44 AUs defined by the FACS system. Besides, in many instances strong assumptions are made to make the problem more tractable (e.g., images contain faces with no facial hair or glasses, the illumination is constant, the subjects are young and of the same ethnicity). Only the method of Braathen et al. (2002) deals with rigid head motions, and only the method of Essa and Pentland (1997) can handle distractions like facial hair (i.e., beard, moustache) and glasses. None of the automated facial expression analyzers proposed in the literature to date fills in missing parts of the observed face; that is, none perceives a whole face when a part of it is occluded (i.e., by a hand or some other object). Also, though

the conclusions generated by an automated facial expression analyzer are affected by input data certainty, robustness of the applied processing mechanisms, and so forth, except for the system proposed by Pantic and Rothkrantz (2004), no existing system for automatic facial expression analysis calculates the output data certainty.

In spite of repeated references to the need for a readily accessible reference set of static images and image sequences of faces that could provide a basis for benchmarks for efforts in automating FACS coding, no database of images exists that is shared by all diverse facial-expression-research communities. In general, only isolated pieces of such a facial database exist. An example is the unpublished database of Ekman-Hager Facial Action Exemplars. It has been used by Bartlett et al. (1999), Donato et al. (1999), and Tian et al. (2001) to train and test their methods for AU detection from face image sequences. The facial database made publicly available, but still not used by all diverse facial-expression-research communities, is the Cohn-Kanade AU-coded Face Expression Image Database (Kanade et al., 2000). None of these databases contains images of faces in profile view, none contains images of all possible single-AU activations, and none contains images of spontaneous facial expressions. Also, the metadata associated with each database object usually does not identify the temporal segments of AUs shown in the face video in question. This lack of suitable and common training and testing material forms the major impediment to comparing, resolving, and extending the issues concerned with facial micro-action detection from face video. It is, therefore, a critical issue that should be addressed in the nearest possible future.

CONCLUSION

Faces are tangible projector panels of the mechanisms that govern our emotional and social behaviors. Analysis of facial expressions in terms of rapid facial signals (i.e., in terms of the activity of the facial muscles causing the visible changes in facial expression) is, therefore, a highly intriguing problem. While the automation of the entire process of facial action coding from digitized images would be enormously beneficial for fields as diverse as medicine, law,

communication, education, and computing, we should recognize the likelihood that such a goal still belongs to the future. The critical issues concern the establishment of basic understanding of how to achieve automatic spatio-temporal facial-gesture analysis from multiple views of the human face and the establishment of a readily accessible centralized repository of face images that could provide a basis for benchmarks for efforts in the field.

REFERENCES

- Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, *36*, 253-263.
- Braathen, B., Bartlett, M.S., Littlewort, G., Smith, E., & Movellan, J.R. (2002). An approach to automatic recognition of spontaneous facial actions. *Proceedings of the International Conference on Face and Gesture Recognition (FGR'02)*.
- Cohn, J.F., Zlochower, A.J., Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual faces coding. *Psychophysiology*, *36*, 35-43.
- Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J. (1999). Classifying facial actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *21*(10), 974-989.
- Ekman, P., & Friesen, W.V. (1978). *Facial action coding system*. Palo Alto, CA: Consulting Psychologist Press.
- Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial action coding system*. Salt Lake City, UT: Human Face.
- Essa, I., & Pentland, A. (1997). Coding, analysis, interpretation and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *19*(7), 757-763.
- Kanade, T., Cohn, J., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings of the International Conference on Face and Gesture Recognition*.

Lewis, M., & Haviland-Jones, J.M. (Eds.). (2000). *Handbook of emotions*. New York, NY: Guilford Press.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions*, *E74*(10), 3474-3483.

Mendolia, M., & Kleck, R.E. (1991). Watching people talk about their emotions—Inferences in response to full-face vs. profile expressions. *Motivation and Emotion* *15*(4), 229-242.

Mitra, S., & Liu, Y. (2004). Local facial asymmetry for expression classification. *Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

Pantic, M., & Patras, I. (2004). Temporal modeling of facial actions from face profile image sequences. *Proceedings of the International Conference on Multimedia and Expo*.

Pantic, M., & Rothkrantz, L.J.M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *IEEE*, *91*(9), 1370-1390.

Pantic, M., & Rothkrantz, L.J.M. (2004). Facial action recognition for facial expression analysis from static face images. *IEEE Trans. Systems, Man, and Cybernetics – Part B*, *34*(3), 1449-1461.

Scherer, K.R., & Ekman, P. (Eds.). (1982). *Handbook of methods in non-verbal behavior research*. Cambridge, MA: Cambridge University Press.

Schmidt, K.L., & Cohn, J.F. (2001). Dynamics of facial expression: Normative characteristics and individual differences. *Proceedings of the International Conference on Multimedia and Expo*.

Tian, Y., Kanade, T., & Cohn, J.F. (2001). Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *23*(2), 97-115.

Valstar, M.F., Patras, I., & Pantic, M. (2004). Facial action unit recognition using temporal templates. *Proceedings of the International Workshop on Robot-Human Interaction*.

KEY TERMS

Ambient Intelligence: The merging of mobile communications and sensing technologies with the aim of enabling a pervasive and unobtrusive intelligence in the surrounding environment supporting the activities and interactions of the users. Technologies like face-based interfaces and affective computing are inherent ambient-intelligence technologies.

Automatic Facial Expression Analysis: A process of locating the face in an input image, extracting facial features from the detected face region, and classifying these data into some facial-expression-interpretative categories such as facial muscle action categories, emotion (affect), attitude, and so forth.

Face-Based Interface: Regulating (at least partially) the command flow that streams between the user and the computer by means of facial signals. This means associating certain commands (e.g., mouse pointing, mouse clicking, etc.) with certain facial signals (e.g., gaze direction, winking, etc.). Face-based interface can be effectively used to free computer users from classic keyboard and mouse commands.

Face Synthesis: A process of creating a talking head that is able to speak, display (appropriate) lip movements during speech, and display expressive facial movements.

Lip Reading: The human ability to hear in noisy environments by analyzing visible speech signals; that is, by analyzing the movements of the lips and the surrounding facial region. Integrating both visual speech processing and acoustic speech processing results in a more robust bimodal (audiovisual) speech processing.

Machine Learning: A field of computer science concerned with the question of how to construct computer programs that automatically improve with experience. The key algorithms that form the core of machine learning include neural networks, genetic algorithms, support vector machines, Bayesian networks, and Markov models.

Machine Vision: A field of computer science concerned with the question of how to construct computer programs that automatically analyze images and produce descriptions of what is imaged.