20

Machine Analysis of Facial Expressions

Maja Pantic [1] and Marian Stewart Bartlett [2]
[1] *Computing Department, Imperial College London,*
[2] *Inst. Neural Computation, University of California*
[1] *UK,* [2] *USA*

1. Human Face and Its Expression

The human face is the site for major sensory inputs and major communicative outputs. It houses the majority of our sensory apparatus as well as our speech production apparatus. It is used to identify other members of our species, to gather information about age, gender, attractiveness, and personality, and to regulate conversation by gazing or nodding. Moreover, the human face is our preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression (Keltner & Ekman, 2000). Thus, the human face is a multi-signal input-output communicative system capable of tremendous flexibility and specificity (Ekman & Friesen, 1975). In general, the human face conveys information via four kinds of signals.

(a) *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 contribute to an individual's appearance and are usually exploited for person identification.

(b) *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. Note that these signals might diminish the distinctness of the boundaries of the facial features and impede recognition of the rapid facial signals.

(c) *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. Note that these signals might obscure facial features or, conversely, might enhance them.

(d) *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*.

All four classes of signals contribute to person identification, gender recognition, attractiveness assessment, and personality prediction. In Aristotle's time, a theory was proposed about mutual dependency between static facial signals (physiognomy) and personality: "soft hair reveals a coward, strong chin a stubborn person, and a smile a happy person". Today, few psychologists share the belief about the meaning of soft hair and strong chin, but many believe that rapid facial signals (facial expressions) communicate emotions (Ekman & Friesen, 1975; Ambady & Rosenthal, 1992; Keltner & Ekman, 2000) and personality traits (Ambady & Rosenthal, 1992). More specifically, types of messages communicated by rapid facial signals include the following (Ekman & Friesen, 1969; Pantic et al., 2006):

(a) affective / attitudinal states and moods,1 e.g., joy, fear, disbelief, interest, dislike, stress,
(b) emblems, i.e., culture-specific communicators like wink,
(c) manipulators, i.e., self-manipulative actions like lip biting and yawns,
(d) illustrators, i.e., actions accompanying speech such as eyebrow flashes,
(e) regulators, i.e., conversational mediators such as the exchange of a look, head nodes and smiles.

1.1 Applications of Facial Expression Measurement Technology

Given the significant role of the face in our emotional and social lives, it is not surprising that the potential benefits from efforts to automate the analysis of facial signals, in particular rapid facial signals, are varied and numerous (Ekman et al., 1993), especially when it comes to computer science and technologies brought to bear on these issues (Pantic, 2006).

As far as natural 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 sub-systems, including automated tools for tracking gaze and focus of attention, lip reading, bimodal speech processing, face / visual speech synthesis, and face-based command issuing. 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, recognizing the person's facial signals and making the avatar respond to those using synthesized speech and facial expressions is important. Combining facial expression spotting with facial expression interpretation in terms of labels like "did not understand", "disagree", "inattentive", and "approves" could be employed as a tool for monitoring human reactions during videoconferences, web-based lectures, and automated tutoring sessions. Attendees' facial expressions will inform the speaker (teacher) of the need to adjust the (instructional) presentation.

The focus of the relatively recently initiated research area of *affective computing* lies on sensing, detecting and interpreting human affective states and devising appropriate means for handling this affective information in order to enhance current HCI designs (Picard, 1997). The tacit assumption is that in many situations human-machine interaction could be improved by the introduction of machines that can adapt to their users (think about computer-based advisors, virtual information desks, on-board computers and navigation systems, pacemakers, etc.). The information about when the existing processing should be

[1] In contrast to traditional approach, which lists only (basic) emotions as the first type of messages conveyed by rapid facial signals (Ekman & Friesen, 1969), we treat this type of messages as being correlated not only to emotions but to other attitudinal states, social signals, and moods as well. We do so becuase cues identifying attitudinal states like interest and boredom, to those underlying moods, and to those disclosing social signaling like empathy and antipathy are all visualy detectable from someone's facial expressions (Pantic et al., 2005, 2006).

adapted, the importance of such an adaptation, and how the processing/reasoning should be adapted, involves information about the how the user feels (e.g. confused, irritated, frustrated, interested). As facial expressions are our direct, naturally preeminent means of communicating emotions, machine analysis of facial expressions forms an indispensable part of affective HCI designs (Pantic & Rothkrantz, 2003; Maat & Pantic, 2006).

Automatic assessment of boredom, fatigue, and stress, will be highly valuable in situations where firm attention to a crucial but perhaps tedious task is essential, such as aircraft and air traffic control, space flight and nuclear plant surveillance, or simply driving a ground vehicle like a truck, train, or car. If these negative affective states could be detected in a timely and unobtrusive manner, appropriate alerts could be provided, preventing many accidents from happening. Automated detectors of fatigue, depression and anxiety could form another step toward personal wellness technologies. Automating such assessment becomes increasingly important in an aging population to prevent medical practitioners from becoming overburdened.

Monitoring and interpreting facial signals can also provide important information to lawyers, police, security, and intelligence agents regarding deception and attitude. Automated facial reaction monitoring could form a valuable tool in law enforcement, as now only informal interpretations are typically used. Systems that can recognize friendly faces or, more importantly, recognize unfriendly or aggressive faces and inform the appropriate authorities represent another application of facial measurement technology.

1.2 Outline of the Chapter

This chapter introduces recent advances in machine analysis of facial expressions. It first surveys the problem domain, describes the problem space, and examines the state of the art. Then it describes several techniques used for automatic facial expression analysis that were recently proposed by the authors. Four areas will receive particular attention: face detection, facial feature extraction, facial muscle action detection, and emotion recognition. Finally, some of the scientific and engineering challenges are discussed and recommendations for achieving a better facial expression measurement technology are outlined.

2. Automatic Facial Expression Analysis: Problem Space and State of the Art

Because of its practical importance explained above and the theoretical interest of cognitive and medical scientists (Ekman et al., 1993; Young, 1998; Cohen, 2006), machine analysis of facial expressions attracted the interest of many researchers. However, although humans detect and analyze faces and facial expressions in a scene with little or no effort, development of an automated system that accomplishes this task is rather difficult.

2.1 Level of Description: Action Units and Emotions

Two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) detection and facial muscle action (action unit) detection. For exhaustive surveys of the related work, readers are referred to: Samal & Iyengar (1992) for an overview of early works, Tian et al. (2005) and Pantic (2006) for surveys of techniques for detecting facial muscle actions, and Pantic and Rothkrantz (2000, 2003) for surveys of facial affect recognition methods.

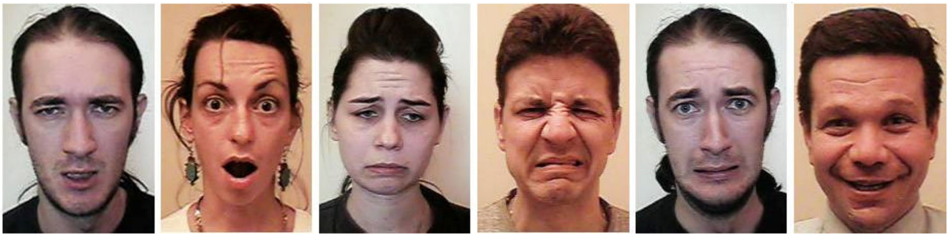

Figure 1. Prototypic facial expressions of six basic emotions: anger, surprise, sadness, disgust, fear, and happiness

These two streams stem directly from two major approaches to facial expression measurement in psychological research (Cohn, 2006): message and sign judgment. The aim of message judgment is to *infer* what underlies a displayed facial expression, such as affect or personality, while the aim of sign judgment is to *describe* the "surface" of the shown behavior, such as facial movement or facial component shape. Thus, a brow furrow can be judged as "anger" in a message-judgment and as a facial movement that lowers and pulls the eyebrows closer together in a sign-judgment approach. While message judgment is all about interpretation, sign judgment attempts to be objective, leaving inference about the conveyed message to higher order decision making.

As indicated by Cohn (2006), most commonly used facial expression descriptors in message judgment approaches are the six basic emotions (fear, sadness, happiness, anger, disgust, surprise; see Figure 1), proposed by Ekman and discrete emotion theorists, who suggest that these emotions are universally displayed and recognized from facial expressions (Keltner & Ekman, 2000). This trend can also be found in the field of automatic facial expression analysis. Most facial expressions analyzers developed so far target human facial affect analysis and attempt to recognize a small set of prototypic emotional facial expressions like happiness and anger (Pantic et al., 2005a). Automatic detection of the six basic emotions in posed, controlled displays can be done with reasonably high accuracy. However detecting these facial expressions in the less constrained environments of real applications is a much more challenging problem which is just beginning to be explored. There have also been a few tentative efforts to detect cognitive and psychological states like interest (El Kaliouby & Robinson, 2004), pain (Bartlett et al., 2006), and fatigue (Gu & Ji, 2005).

In sign judgment approaches (Cohn & Ekman, 2005), a widely used method for manual labeling of facial actions is the Facial Action Coding System (FACS; Ekman & Friesen, 1978, Ekman et al., 2002). FACS associates facial expression changes with actions of the muscles that produce them. It defines 44 different action units (AUs), which are considered to be the smallest visually discernable facial movements (e.g, see Figure 2). FACS also provides the rules for recognition of AUs' temporal segments (onset, apex and offset) in a face video. Using FACS, human coders can manually code nearly any anatomically possible facial display, decomposing it into the AUs and their temporal segments that produced the display. As AUs are independent of interpretation, they can be used for any higher order decision making process including recognition of basic emotions (Ekman et al., 2002), cognitive states like (dis)agreement and puzzlement (Cunningham et al., 2004), psychological states like suicidal depression (Heller & Haynal, 1997) or pain (Williams, 2002; Craig et al., 1991), and social signals like emblems (i.e., culture-specific interactive signals like wink), regulators (i.e., conversational mediators like nod and smile), and illustrators (i.e., cues accompanying speech like raised eyebrows) (Ekman & Friesen, 1969). Hence, AUs are very suitable to be used as mid-level parameters in automatic facial behavior analysis, as the thousands of anatomically possible expressions (Cohn & Ekman, 2005) can be described as combinations of 5 dozens of AUs and can be mapped to any higher order facial display interpretation.

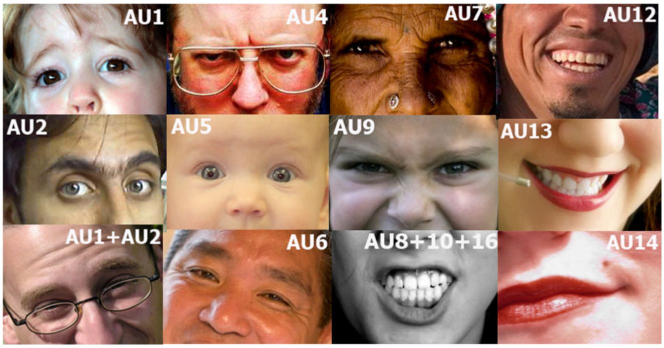

Figure 2(a). Examples of facial action units (AUs) and their combinations defined in FACS

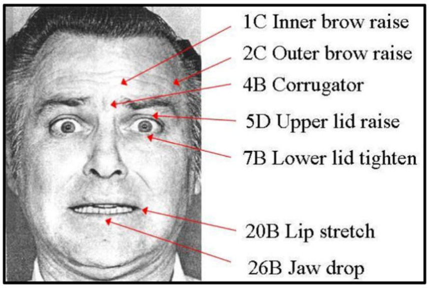

Figure 2(b). Example FACS codes for a prototypical expression of fear. FACS provides a 5-point intensity scale (A-E) to describe AU intensity variation; e.g., 26B stands for a weak jaw drop

FACS provides an objective and comprehensive language for describing facial expressions and relating them back to what is known about their meaning from the behavioral science literature. Because it is comprehensive, FACS also allows for the discovery of new patterns related to emotional or situational states. For example, what are the facial behaviors associated with driver fatigue? What are the facial behaviors associated with states that are critical for automated tutoring systems, such as interest, boredom, confusion, or comprehension? Without an objective facial measurement system, we have a chicken- and-egg problem. How do we build systems to detect comprehension, for example, when we don’t know for certain what faces do when students are comprehending? Having subjects pose states such as comprehension and confusion is of limited use since there is a great deal of evidence that people do different things with their faces when posing versus during a spontaneous experience (Ekman, 1991, 2003). Likewise, subjective labeling of expressions has also been shown to be less reliable than objective coding for finding relationships between facial expression and other state variables. Some examples of this include the failure of subjective labels to show associations between smiling and other measures of happiness, and it was not until FACS coding was introduced that a strong relationship was found, namely that expressions containing an eye region movement in addition to the mouth movement (AU12+6) were correlated with happiness, but expressions just containing the mouth smile (AU12) did not (Ekman, 2003). Another example where subjective judgments of expression failed to find relationships which were later found with FACS is the failure of naive subjects to differentiate deception and intoxication from facial display, whereas reliable differences were shown with FACS (Sayette et al., 1992). Research based upon FACS has also shown that facial actions can show differences between those telling the truth and lying at a much higher accuracy level than naive subjects making subjective judgments of the same faces (Frank & Ekman, 2004).

Objective coding with FACS is one approach to the problem of developing detectors for state variables such as comprehension and confusion, although not the only one. Machine learning of classifiers from a database of spontaneous examples of subjects in these states is another viable approach, although this carries with it issues of eliciting the state, and assessment of whether and to what degree the subject is experiencing the desired state. Experiments using FACS face the same challenge, although computer scientists can take advantage of a large body of literature in which this has already been done by behavioral scientists. Once a database exists, however, in which a state has been elicited, machine learning can be applied either directly to image primitives, or to facial action codes. It is an open question whether intermediate representations such as FACS are the best approach to recognition, and such questions can begin to be addressed with databases such as the ones described in this chapter. Regardless of which approach is more effective, FACS provides a general purpose representation that can be useful for many applications. It would be time consuming to collect a new database and train application-specific detectors directly from image primitives for each new application. The speech recognition community has converged on a strategy that combines intermediate representations from phoneme detectors plus context-dependent features trained directly from the signal primitives, and perhaps a similar strategy will be effective for automatic facial expression recognition.

It is not surprising, therefore, that automatic AU coding in face images and face image sequences attracted the interest of computer vision researchers. Historically, the first attempts to encode AUs in images of faces in an automatic way were reported by Bartlett et al. (1996), Lien et al. (1998), and Pantic et al. (1998). These three research groups are still the forerunners in this research field. The focus of the research efforts in the field was first on automatic recognition of AUs in either static face images or face image sequences picturing facial expressions produced on command. Several promising prototype systems were reported that can recognize deliberately produced AUs in either (near-) frontal view face images (Bartlett et al., 1999; Tian et al., 2001; Pantic & Rothkrantz, 2004a) or profile view face images (Pantic & Rothkrantz, 2004a; Pantic & Patras, 2006). These systems employ different approaches including expert rules and machine learning methods such as neural networks, and use either feature-based image representations (i.e., use geometric features like facial points; see section 5) or appearance-based image representations (i.e., use texture of the facial skin including wrinkles and furrows; see section 6).

One of the main criticisms that these works received from both cognitive and computer scientists, is that the methods are not applicable in real-life situations, where subtle changes in facial expression typify the displayed facial behavior rather than the exaggerated changes that typify posed expressions. Hence, the focus of the research in the field started to shift to automatic AU recognition in spontaneous facial expressions (produced in a reflex-like manner). Several works have recently emerged on machine analysis of AUs in spontaneous facial expression data (Cohn et al., 2004; Bartlett et al., 2003, 2005, 2006; Valstar et al., 2006). These methods employ probabilistic, statistical, and ensemble learning techniques, which seem to be particularly suitable for automatic AU recognition from face image sequences (Tian et al., 2005; Bartlett et al., 2006).

2.2 Posed vs. Spontaneous Facial Displays

The importance of making a clear distinction between spontaneous and deliberately displayed facial behavior for developing and testing computer vision systems becomes apparent when we examine the neurological substrate for facial expression. There are two distinct neural pathways that mediate facial expressions, each one originating in a different area of the brain. Volitional facial movements originate in the cortical motor strip, whereas the more involuntary, emotional facial actions, originate in the subcortical areas of the brain (e.g. Meihlke, 1973). Research documenting these differences was sufficiently reliable to become the primary diagnostic criteria for certain brain lesions prior to modern imaging methods (e.g. Brodal, 1981.) The facial expressions mediated by these two pathways have differences both in which facial muscles are moved and in their dynamics (Ekman, 1991; Ekman & Rosenberg, 2005). Subcortically initiated facial expressions (the involuntary group) are characterized by synchronized, smooth, symmetrical, consistent, and reflex-like facial muscle movements whereas cortically initiated facial expressions are subject to volitional real-time control and tend to be less smooth, with more variable dynamics (Rinn, 1984; Ekman & Rosenberg, 2005). However, precise characterization of spontaneous expression dynamics has been slowed down by the need to use non-invasive technologies (e.g. video), and the difficulty of manually coding expression intensity frame-by-frame. Thus the importance of video based automatic coding systems.

Furthermore, the two pathways appear to correspond to the distinction between biologically driven versus socially learned facial behavior (Bartlett et al., 2006). Researchers agree, for the most part, that most types of facial expressions are learned like language, displayed under conscious control, and have culturally specific meanings that rely on context for proper interpretation (Ekman, 1989). Thus, the same lowered eyebrow expression that would convey “uncertainty” in North America might convey ”no” in Borneo (Darwin, 1872/1998). On the other hand, there are a limited number of distinct facial expressions of emotion that appear to be biologically wired, produced involuntarily, and whose meanings are similar across all cultures; for example, anger, contempt, disgust, fear, happiness, sadness, and surprise (see section 2.1). There are also spontaneous facial movements that accompany speech. These movements are smooth and ballistic, and are more typical of the subcortical system associated with spontaneous expressions (e.g. Rinn, 1984). There is some evidence that arm-reaching movements transfer from one motor system when they require planning to another when they become automatic, with different dynamic characteristics between the two (Torres & Anderson, 2006). It is unknown whether the same thing happens with learned facial displays. An automated system would enable exploration of such research questions.

As already mentioned above, few works have been recently reported on machine analysis of spontaneous facial expression data (Cohn et al., 2004; Bartlett et al., 2003, 2005, 2006; Valstar et al., 2006). Except of the method for discerning genuine from fake facial expressions of pain described in section 7.3, the only reported effort to automatically discern spontaneous from deliberately displayed facial behavior is that of Valstar et al. (2006). It concerns an automated system for distinguishing posed from spontaneous brow actions (i.e. AU1, AU2, AU4, and their combinations). Conforming with the research findings in psychology, the system was built around characteristics of temporal dynamics of brow actions and employs parameters like speed, intensity, duration, and the occurrence order of brow actions to classify brow actions present in a video as either deliberate or spontaneous facial actions.

2.3 Facial Expression Configuration and Dynamics

Automatic recognition of facial expression configuration (in terms of AUs constituting the observed expression) has been the main focus of the research efforts in the field. However, both the configuration and the dynamics of facial expressions (i.e., the timing and the duration of various AUs) are important for interpretation of human facial behavior. The body of research in cognitive sciences, which argues that the dynamics of facial expressions are crucial for the interpretation of the observed behavior, is ever growing (Basilli, 1978; Russell & Fernandez-Dols, 1997; Ekman & Rosenberg, 2005; Ambadar et al., 2005). Facial expression temporal dynamics are essential for categorization of complex psychological states like various types of pain and mood (Williams, 2002). They represent a critical factor for interpretation of social behaviors like social inhibition, embarrassment, amusement, and shame (Keltner, 1997; Costa t al., 2001). They are also a key parameter in differentiation between posed and spontaneous facial displays (Ekman & Rosenberg, 2005). For instance, spontaneous smiles are smaller in amplitude, longer in total duration, and slower in onset and offset time than posed smiles (e.g., a polite smile) (Ekman, 2003). Another study showed that spontaneous smiles, in contrast to posed smiles, can have multiple apexes (multiple rises of the mouth corners – AU12) and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1s (Cohn & Schmidt, 2004). Similarly, it has been shown that the differences between spontaneous and deliberately displayed brow actions (AU1, AU2, AU4) is in the duration and the speed of onset and offset of the actions and in the order and the timing of actions' occurrences (Valstar et al. 2006).

In spite of these findings, the vast majority of the past work in the field does not take dynamics of facial expressions into account when analyzing shown facial behavior. Some of the past work in the field has used aspects of temporal dynamics of facial expression such as the speed of a facial point displacement or the persistence of facial parameters over time (e.g., Zhang & Ji, 2005; Tong et al., 2006; Littlewort et al., 2006). However, only three recent studies analyze explicitly the temporal dynamics of facial expressions. These studies explore automatic segmentation of AU activation into temporal segments (neutral, onset, apex, offset) in frontal- (Pantic & Patras, 2005; Valstar & Pantic, 2006a) and profile-view (Pantic & Patras, 2006) face videos. The works of Pantic & Patras (2005, 2006) employ rule-based reasoning to encode AUs and their temporal segments. In contrast to biologically inspired learning techniques (such as neural networks), which emulate human unconscious problem solving processes, rule-based techniques are inspired by human conscious problem solving processes. However, studies in cognitive sciences, like the one on "thin slices of behavior" (Ambady & Rosenthal, 1992), suggest that facial displays are neither encoded nor decoded at an intentional, conscious level of awareness. They may be fleeting changes in facial appearance that we still accurately judge in terms of emotions or personality even from very brief observations. In turn, this finding suggests that learning techniques inspired by human unconscious problem solving may be more suitable for facial expression recognition than those inspired by human conscious problem solving (Pantic et al., 2005a). Experimental evidence supporting this assumption for the case of prototypic emotional facial expressions was recently reported (Valstar & Pantic, 2006b). Valstar & Pantic (2006a) also presented experimental evidence supporting this assumption for the case of expression configuration detection and its temporal activation model (neutral → onset → apex → offset) recognition.

2.4 Facial Expression Intensity, Intentionality and Context Dependency

Facial expressions can vary in intensity. By intensity we mean the relative degree of change in facial expression as compared to a relaxed, neutral facial expression. In the case of a smile, for example, the intensity of the expression can be characterized as the degree of upward and outward movement of the mouth corners, that is, as the degree of perceivable activity in the Zygomaticus Major muscle (AU12) away from its resting, relaxed state (Duchenne, 1862/1990; Ekman & Friesen, 1978). It has been experimentally shown that the expression decoding accuracy and the perceived intensity of the underlying affective state vary linearly with the physical intensity of the facial display (Hess et al., 1997). Hence, explicit analysis of expression intensity variation is very important for accurate expression interpretation, and is also essential to the ability to distinguish between spontaneous and posed facial behavior discussed in the previous sections. While FACS provides a 5-point intensity scale to describe AU intensity variation and enable manual quantification of AU intensity (Ekman et al. 2002; Figure 2(b)), fully automated methods that accomplish this task are yet to be developed. However, first steps toward this goal have been made. Some researchers described changes in facial expression that could be used to represent intensity variation automatically (Essa & Pentland, 1997; Kimura & Yachida, 1997; Lien et al., 1998), and an effort toward implicit encoding of intensity was reported by Zhang & Ji (2005). Automatic coding of intensity variation was explicitly compared to manual coding in Bartlett et al. (2003a; 2006). They found that the distance to the separating hyperplane in their learned classifiers correlated significantly with the intensity scores provided by expert FACS coders.

Rapid facial signals do not usually convey exclusively one type of messages but may convey any of the types (e.g., blinking is usually a manipulator but it may be displayed in an expression of confusion). It is crucial to determine which type of message a shown facial expression communicates since this influences the interpretation of it (Pantic & Rothkrantz, 2003). For instance, squinted eyes may be interpreted as sensitivity of the eyes to bright light if this action is a reflex (a manipulator), as an expression of disliking if this action has been displayed when seeing someone passing by (affective cue), or as an illustrator of friendly anger on friendly teasing if this action has been posed (in contrast to being unintentionally displayed) during a chat with a friend, to mention just a few possibilities. To interpret an observed facial signal, it is important to know the context in which the observed signal has been displayed – where the expresser is (outside, inside, in the car, in the kitchen, etc.), what his or her current task is, are other people involved, and who the expresser is. Knowing the expresser is particularly important as individuals often have characteristic facial expressions and may differ in the way certain states (other than the basic emotions) are expressed. Since the problem of context-sensing is extremely difficult to solve (if possible at all) for a general case, pragmatic approaches (e.g., activity/application- and user-centered approach) should be taken when learning the grammar of human facial behavior (Pantic et al., 2005a, 2006). However, except for a few works on user-profiled interpretation of facial expressions like those of Fasel et al. (2004) and Pantic & Rothkrantz (2004b), virtually all existing automated facial expression analyzers are context insensitive. Although machine-context sensing, that is, answering questions like who is the user, where is he or she, and what is he or she doing, has witnessed recently a number of significant advances (Nock et al., 2004, Pantic et al. 2006), the complexity of this problem makes context-sensitive facial expression analysis a significant research challenge.

2.5 Facial Expression Databases and Ground Truth

To develop and evaluate facial behavior analyzers capable of dealing with different dimensions of the problem space as defined above, large collections of training and test data are needed (Pantic & Rothkrantz, 2003; Pantic et al., 2005a; Tian et al., 2005; Bartlett et al., 2006).

Picard (1997) outlined five factors that influence affective data collection:

(a) Spontaneous versus posed: Is the emotion elicited by a situation or stimulus that is outside the subject's control or the subject is asked to elicit the emotion?
(b) Lab setting versus real-world: Is the data recording taking place in a lab or the emotion is recorded in the usual environment of the subject?
(c) Expression versus feeling: Is the emphasis on external expression or on internal feeling?
(d) Open recording versus hidden recording: Is the subject aware that he is being recorded?
(e) Emotion-purpose versus other-purpose: Does the subject know that he is a part of an experiment and the experiment is about emotion?

A complete overview of existing, publicly available datasets that can be used in research on automatic facial expression analysis is given by Pantic et al. (2005b). In general, there is no comprehensive reference set of face images that could provide a basis for all different efforts in the research on machine analysis of facial expressions. Only isolated pieces of such a facial database exist. An example is the unpublished database of Ekman-Hager Facial Action Exemplars (Ekman et al., 1999). It has been used by several research groups (e.g., Bartlett et al., 1999; Tian et al., 2001) to train and test their methods for AU detection from frontal-view facial expression sequences. Another example is JAFFE database (Lyons et al., 1999), which contains in total 219 static images of 10 Japanese females displaying posed expressions of six basic emotions and was used for training and testing various existing methods for recognition of prototypic facial expressions of emotions (Pantic et al., 2003). An important recent contribution to the field is the Yin Facial Expression Database (Yin et al., 2006), which contains 3D range data for prototypical expressions at a variety of intensities.

The Cohn-Kanade facial expression database (Kanade et al., 2000) is the most widely used database in research on automated facial expression analysis (Tian et al., 2005; Pantic et al., 2005a). This database contains image sequences of approximately 100 subjects posing a set of 23 facial displays, and contains FACS codes in addition to basic emotion labels. The release of this database to the research community enabled a large amount of research on facial expression recognition and feature tracking. Two main limitations of this facial expression data set are as follows. First, each recording ends at the apex of the shown expression, which limits research of facial expression temporal activation patterns (onset → apex → offset). Second, many recordings contain the date/time stamp recorded over the chin of the subject. This makes changes in the appearance of the chin less visible and motions of the chin difficult to track.

To fill this gap, the MMI facial expression database was developed (Pantic et al., 2005b). It has two parts: a part containing deliberately displayed facial expressions and a part containing spontaneous facial displays. The first part contains over 4000 videos as well as over 600 static images depicting facial expressions of single AU activation, multiple AU activations, and six basic emotions. It has profile as well as frontal views, and was FACS coded by two certified coders. The second part of the MMI facial expression database contains currently 65 videos of spontaneous facial displays, that were coded in terms of displayed AUs and emotions by two certified coders. Subjects were 18 adults 21 to 45 years old and 11 children 9 to 13 years old; 48% female, 66% Caucasian, 30% Asian and 4% African. The recordings of 11 children were obtained during the preparation of a Dutch TV program, when children were told jokes by a professional comedian or were told to mimic how they would laugh when something is not funny. The recordings contain mostly facial expressions of different kinds of laughter and were made in a TV studio, using a uniform background and constant lighting conditions. The recordings of 18 adults were made in subjects' usual environments (e.g., home), where they were shown segments from comedies, horror movies, and fear-factor series. The recordings contain mostly facial expressions of different kinds of laughter, surprise, and disgust expressions, which were accompanied by (often large) head motions, and were made under variable lighting conditions. Although the MMI facial expression database is the most comprehensive database for research on automated facial expression analysis, it still lacks metadata for the majority of recordings when it comes to frame-based AU coding. Further, although the MMI database is probably the only publicly available dataset containing recordings of spontaneous facial behavior at present, it still lacks metadata about the context in which these recordings were made such the utilized stimuli, the environment in which the recordings were made, the presence of other people, etc.

Another database of spontaneous facial expressions was collected at UT Dallas (O'Toole et al., 2005). Similarly to the second part of the MMI facial expression database, facial displays were elicited using film clips. In the case of the UT Dallas database, however, there is no concurrent measure of expression content beyond the stimulus category. Yet, since subjects often do not experience the intended emotion and sometimes experience another one (e.g., disgust or annoyance instead of humor), concurrent measure of expression content beyond the stimulus category is needed. In other words, as in the case of the second part of the MMI facial expression database, coding in terms of displayed AUs and emotions independently of the stimulus category is needed.

Mark Frank, in collaboration with Javier Movellan and Marian Bartlett, has collected a dataset of spontaneous facial behavior in an interview paradigm with rigorous FACS coding (Bartlett et al. 2006). This datased, called the RU-FACS Spontaneous Expression Dataset, consists of 100 subjects participating in a 'false opinion' paradigm. In this paradigm, subjects first fill out a questionnaire regarding their opinions about a social or political issue. Subjects are then asked to either tell the truth or take the opposite opinion on an issue where they rated strong feelings, and convince an interviewer they are telling the truth. Interviewers were retired police and FBI agents. A high-stakes paradigm was created by giving the subjects $50 if they succeeded in fooling the interviewer, whereas if they were caught they were told they would receive no cash, and would have to fill out a long and boring questionnaire. In practice, everyone received a minimum of $10 for participating, and no one had to fill out the questionnaire. This paradigm has been shown to elicit a wide range of emotional expressions as well as speech-related facial expressions. This dataset is particularly challenging both because of speech-related mouth movements, and also because of out-of-plane head rotations which tend to be present during discourse. Subjects faces were digitized by four synchronized Dragonfly cameras from Point Grey (frontal, two partial profiles at 30 degrees, and one view from below). Two minutes of each subject's behavior is being FACS coded by two certified FACS coders. FACS codes include the apex frame as well as the onset and offset frame for each action unit (AU). To date, 33 subjects have been FACS-coded. This dataset will be made available to the research community once the FACS coding is completed.

With the exception of these problems concerned with acquiring valuable data and the related ground truth, another important issue is how does one construct and administer such a large facial expression benchmark database. Except of the MMI facial expression database (Pantic et al., 2005b), which was built as a web-based direct-manipulation application, allowing easy access and easy search of the available images, the existing facial expression databases are neither easy to access nor easy to search. In general, once the permission for usage is issued, large, unstructured files of material are sent. Other related questions are the following. How does one facilitate reliable, efficient, and secure inclusion of objects constituting this database? How could the performance of a tested automated system be included into the database? How should the relationship between the performance and the database objects used in the evaluation be defined? Pantic et al. (2003, 2005a, 2005b) emphasized a number of specific, research and development efforts needed to address the aforementioned problems. Nonetheless, note that their list of suggestions and recommendations is not exhaustive of worthwhile contributions.

3. Face Detection

The first step in facial information processing is face detection, i.e., identification of all regions in the scene that contain a human face. The problem of *finding faces* should be solved regardless of clutter, occlusions, and variations in head pose and lighting conditions. The presence of non-rigid movements due to facial expression and a high degree of variability in facial size, color and texture make this problem even more difficult. Numerous techniques have been developed for face detection in still images (Yang et al., 2002; Li & Jain, 2005). However, most of them can detect only upright faces in frontal or near-frontal view. The efforts that had the greatest impact on the community (as measured by, e.g., citations) include the following.

Rowley et al. (1998) used a multi-layer neural network to learn the face and non-face patterns from the intensities and spatial relationships of pixels in face and non-face images. Sung and Poggio (1998) proposed a similar method. They used a neural network to find a discriminant function to classify face and non-face patterns using distance measures. Moghaddam and Pentland (1997) developed a probabilistic visual learning method based on density estimation in a high-dimensional space using an eigenspace decomposition. The method was applied to face localization, coding and recognition. Pentland et al. (1994) developed a real-time, view-based and modular (by means of incorporating salient features such as the eyes and the mouth) eigenspace description technique for face recognition in variable pose. Another method that can handle out-of-plane head motions is the statistical method for 3D object detection proposed by Schneiderman and Kanade (2000). Other such methods, which have been recently proposed, include those of Huang and Trivedi (2004) and Wang and Ji (2004). Most of these methods emphasize statistical learning techniques and use appearance features.

Arguably the most commonly employed face detector in automatic facial expression analysis is the real-time face detector proposed by Viola and Jones (2004). This detector consists of a cascade of classifiers trained by AdaBoost. Each classifier employs integral image filters, also called "box filters," which are reminiscent of Haar Basis functions, and can be computed very fast at any location and scale. This is essential to the speed of the detector. For each stage in the cascade, a subset of features is chosen using a feature selection procedure based on AdaBoost.

There are several adapted versions of the Viola-Jones face detector and the one that is employed by the systems discussed in detail in this chapter was proposed by Fasel et al. (2005). It uses GentleBoost instead of AdaBoost. GentleBoost uses the continuous output of each filter rather than binarizing it. A description of Gentle Boost classification can be found in Friedman et al. (2000).

4. Facial Feature Extraction

After the presence of a face has been detected in the observed scene, the next step is to extract the information about the displayed facial signals. The problem of *facial feature extraction* from regions in the scene that contain a human face may be divided into at least three dimensions (Pantic & Rothkrantz, 2000):

(a) Is temporal information used?
(b) Are the features holistic (spanning the whole face) or analytic (spanning subparts of the face)?
(c) Are the features view- or volume based (2D/3D)?

Given this glossary and if the goal is face recognition, i.e., identifying people by looking at their faces, most of the proposed approaches adopt 2D holistic static facial features. On the other hand, many approaches to automatic facial expression analysis adopt 2D analytic spatio-temporal facial features (Pantic & Rothkrantz, 2003). This finding is also consistent with findings from psychological research suggesting that the brain processes faces holistically rather than locally whilst it processes facial expressions locally (Bassili, 1978). What is, however, not entirely clear yet is whether information on facial expression is passed to the identification process to aid person recognition or not. Some experimental data suggest this (Martinez, 2003; Roark et al., 2003). For surveys of computer vision efforts aimed at face recognition, the readers are referred to: Zhao et al. (2003), Bowyer (2004), and Li and Jain (2005).

Most of the existing facial expression analyzers are directed toward 2D spatiotemporal facial feature extraction, including the methods proposed by the authors and their respective research teams. The usually extracted facial features are either *geometric features* such as the shapes of the facial components (eyes, mouth, etc.) and the locations of facial fiducial points (corners of the eyes, mouth, etc.) or *appearance features* representing the texture of the facial skin including wrinkles, bulges, and furrows. Typical examples of geometric-feature-based methods are those of Gokturk et al. (2002), who used 19 point face mesh, of Chang et al. (2006), who used a shape model defined by 58 facial landmarks, and of Pantic and her collegues (Pantic & Rothkrantz, 2004; Pantic & Patras, 2006; Valstar & Pantic, 2006a), who used a set of facial characteristic points like the ones illustrated in Figure 3. Typical examples of *hybrid*, geometric- and appearance-feature-based methods are those of Tian et al. (2001), who used shape-based models of eyes, eyebrows and mouth and transient features like crows-feet wrinkles and nasolabial furrow, and of Zhang and Ji (2005), who used 26 facial points around the eyes, eyebrows, and mouth and the same transient features as Tian et al (2001). Typical examples of appearance-feature-based methods are those of Bartlett et al. (1999, 2005, 2006) and Guo and Dyer (2005), who used Gabor wavelets, of Anderson and McOwen (2006), who used a holistic, monochrome, spatial-ratio face template, and of Valstar et al. (2004), who used temporal templates. It has been reported that methods based on geometric features are often outperformed by those based on appearance features using, e.g., Gabor wavelets or eigenfaces (Bartlett et al., 1999). Certainly, this may depend on the classification method and/or machine learning approach which takes the features as input. Recent studies like that of Pantic & Patras (2006), Valstar and Pantic (2006a), and those presented in this chapter, show that in some cases geometric features can outperform appearance-based ones. Yet, it seems that using both geometric and appearance features might be the best choice in the case of certain facial expressions (Pantic & Patras, 2006).

Few approaches to automatic facial expression analysis based on 3D face modelling have been recently proposed. Gokturk et al. (2002) proposed a method for recognition of facial signals like brow flashes and smiles based upon 3D deformations of the face tracked on stereo image streams using a 19-point face mesh and standard optical flow techniques. The work of Cohen et al. (2003) focuses on the design of Bayesian network classifiers for emotion recognition from face video based on facial features tracked by a method called Piecewise Bezier Volume Deformation tracking (Tao & Huang, 1998). This tracker employs an explicit 3D wireframe model consisting of 16 surface patches embedded in Bezier volumes. Cohn et al. (2004) focus on automatic analysis of brow actions and head movements from face video and use a cylindrical head model to estimate the 6 degrees of freedom of head motion (Xiao et al., 2003). Baker and his colleagues developed several algorithms for fitting 2D and combined 2D+3D Active Appearance Models to images of faces (Xiao et al., 2004; Gross et al., 2006), which can be used further for various studies concerning human facial behavior. 3D face modeling is highly relevant to the present goals due to its potential to produce view-independent facial signal recognition systems. The main shortcomings of the current methods concern the need of a large amount of manually annotated training data and an almost always required manual selection of landmark facial points in the first frame of the input video based on which the face model will be warped to fit the face. Automatic facial feature point detection of the kind explained in section 5 offers a solution to these problems.

5. Geometric Facial Feature Extraction and Tracking

5.1 Facial Characteristic Point Detection

Previous methods for facial feature point detection can be classified as either *texture-based methods* (modeling local texture around a given facial point) or *texture- and shape-based methods* (regarding the constellation of all facial points as a shape, which is learned from a set of labeled faces, and trying to fit the shape to any unknown face). A typical texture-based method is that of Holden & Owens (2002), who used log-Gabor wavelets, while a typical texture- and shape-based method is that of Chen et al. (2004), who applied AdaBoost to determine facial feature point candidates for each pixel in an input image and used a shape model as a filter to select the most possible position of feature points.

Although these detectors can be used to localize 20 facial characteristic points illustrated in Figure 3, which are used by the facial expression analyzers developed by Pantic and her team (e.g., Pantic & Patras, 2006; Valstar & Pantic, 2006a), none performs the detection with high accuracy. They usually regard the localization of a point as a SUCCESS if the distance between the automatically labeled point and the manually labeled point is less than 30% of the true inter-ocular distance (the distance between the eyes). However, 30% of the true inter-ocular value is at least 30 pixels in the case of the Cohn-Kanade database samples (Kanade et al., 2000). This means that a bias of 30 pixels for an eye corner would be regarded as SUCCESS even though the width of the whole eye is approximately 60 pixels. This is problematic in the case of facial expression analysis, since subtle changes in the facial expression will be missed due to the errors in facial point localization.

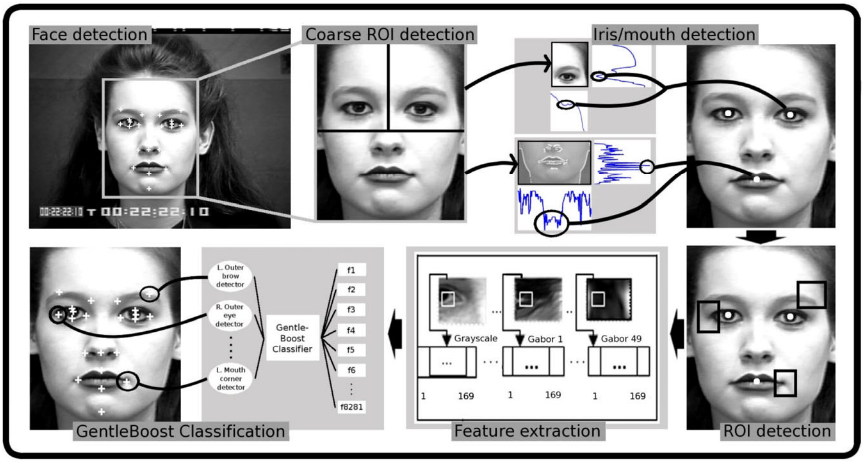

Figure 3. Outline of the fully automated facial point detection method (Vukadinovic & Pantic, 2005)

To handle this, Vukadinovic and Pantic (2005) developed a novel, robust, fully automated facial point detector. The method is illustrated in Figure 3. It is a texture based method – it models local image patches using Gabor wavelets and builds GentleBoost-based point detectors based on these regions. The method operates on the face region detected by the face detector described in section 3. The detected face region is then divided in 20 regions of interest (ROIs), each one corresponding to one facial point to be detected. A combination of heuristic techniques based on the analysis of the vertical and horizontal histograms of the upper and the lower half of the face region image is used for this purpose (Figure 3).

The method uses further individual feature patch templates to detect points in the relevant ROI. These feature models are GentleBoost templates built from both gray level intensities and Gabor wavelet features. Previous work showed that Gabor features were among the most effective texture-based features for face processing tasks (Donato et al., 1999). This finding is also consistent with our experimental data that show the vast majority of features (over 98%) that were selected by the utilized GentleBoost classifier were from the Gabor filter components rather than from the gray level intensities. The essence of the success of Gabor filters is that they remove most of the variability in image due to variation in lighting and contrast, at the same time being robust against small shift and deformation (e.g., Lades et al., 1992; Osadchy et al., 2005). For a thorough analysis of Gabor filters for image representation see (Daugman, 1988).

Feature vector for each facial point is extracted from the 13×13 pixel image patch centered on that point. This feature vector is used to learn the pertinent point's patch template and, in the testing stage, to predict whether the current point represents a certain facial point or not. In total, 13×13×(48+1)=8281 features are used to represent one point (Figure 3). Each feature contains the following information: (i) the position of the pixel inside the 13×13 pixels image patch, (ii) whether the pixel originates from a grayscale or from a Gabor filtered representation of the ROI, and (iii) if appropriate, which Gabor filter has been used (we used a bank of 48 Gabor filters at 8 orientations and 6 spatial frequencies).

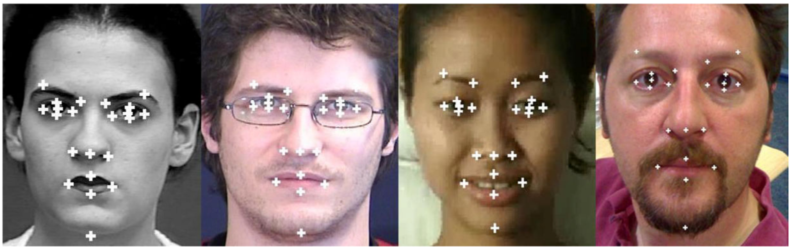

Figure 4. Examples of first-effort results of the facial point detector of Vukadinovic and Pantic (2005) for samples from (left to right): the Cohn-Kanade dataset, the MMI database (posed expressions), the MMI database (spontaneous expressions), and a cell-phone camera

In the training phase, GentleBoost feature templates are learned using a representative set of positive and negative examples. In the testing phase, for a certain facial point, an input 13×13 pixel window (*sliding window*) is slid pixel by pixel across 49 representations of the relevant ROI (grayscale plus 48 Gabor filter representations). For each position of the sliding window, GentleBoost outputs the similarity between the 49-dimensional representation of the sliding window and the learned feature point model. After scanning the entire ROI, the position with the highest similarity is treated as the feature point in question.

Vukadinovic and Pantic trained and tested the facial feature detection method on the first frames of 300 Cohn-Kanade database samples (Kanade et al., 2000), using leave-one-subset-out cross validation. To evaluate the performance of the method, each of the automatically located facial points was compared to the true (manually annotated) point. The authors defined errors with respect to the inter-ocular distance measured in the test image (80 to 120 pixels in the case of image samples from the Cohn-Kanade database). An automatically detected point displaced in any direction, horizontal or vertical, less than 5% of inter-ocular distance (i.e., 4 to 6 pixels in the case of image samples from the Cohn-Kanade database) from the true facial point is regarded as SUCCESS. Overall, an average recognition rate of 93% was achieved for 20 facial feature points using the above described evaluation scheme. Typical results are shown in Figure 4. Virtually all misclassifications (most often encountered with points F1 and M) can be attributed to the lack of consistent rules for manual annotation of the points. For details about this method, see (Vukadinovic & Pantic, 2005).

Fasel and colleagues developed a real-time feature detector using a GentleBoost approach related to the one used for their face detector (Fasel et al., 2005) and combined with a Bayesian model for feature positions (Fasel, 2006). The face is first detected and then the location and scale of the face is used to generate a prior probability distribution for each facial feature. The approach is similar in spirit to that of Vukadinovic and Pantic, but it was trained on 70,000 face snapshots randomly selected from the web. These web images contain greater pose and lighting variation than typical posed expression datasets, and were selected so that the machine learning systems could learn to be robust to such variations, and perform well in the less controlled image conditions of practical applications. When tested on such snapshots, the system obtains a median error of less than 0.05 interocular distance for eye positions, 0.06 for the nose tip, and 0.07 for the mouth center. For the strictly frontal subset of these web snapshots, which still contain substantial lighting variation, median error was 0.04, 0.045, and 0.05 interocular distance for eye, nose, and mouth position. This system could be combined with an approach such as that of Vukadinovic and Pantic to provide more robust initialization for the additional facial feature points.

5.2 Facial Point Tracking

Contractions of facial muscles induce movements of the facial skin and changes in the appearance of facial components such as the eyebrows, nose, and mouth. Since motion of the facial skin produces optical flow in the image, a large number of researchers have studied optical flow tracking (Pantic & Rothkrantz, 2000; 2003; Tian et al., 2005). The optical flow approach to describing face motion has the advantage of not requiring a facial feature extraction stage of processing. Dense flow information is available throughout the entire facial area, regardless of the existence of facial components, even in the areas of smooth texture such as the cheeks and the forehead. Because optical flow is the visible result of movement and is expressed in terms of velocity, it can be used to represent directly facial actions. One of the first efforts to utilize optical flow for recognition of facial expressions was the work of Mase (1991). Thereafter, many other researchers adopted this approach (Pantic & Rothkrantz, 2000; 2003; Tian et al., 2005).

Standard optical flow techniques (e.g., Lucas & Kanade, 1981; Shi & Tomasi, 1994; Barron et al., 1994) are also most commonly used for tracking facial feature points. DeCarlo and Metaxas (1996) presented a model-based tracking algorithm in which a face shape model and motion estimation are integrated using optical flow and edge information. Gokturk et al. (2002) track the points of their 19-point face mesh on the stereo image streams using the standard Lucas-Kanade optical flow algorithm (Lucas & Kanade, 1981). To achieve facial feature point tracking Lien et al. (1998), Tian et al. (2001), and Cohn et al. (2004) used the standard Lucas-Kanade optical flow algorithm too. To realize fitting of 2D and combined 2D+3D Active Appearance Models to images of faces, Xiao et al. (2004) use an algorithm based on an "inverse compositional" extension to the Lucas-Kanade algorithm.

To address the limitations inherent in optical flow techniques such as the accumulation of error and the sensitivity to noise, occlusion, clutter, and changes in illumination, several researchers used sequential state estimation techniques to track facial feature points in image sequences. Both, Zhang and Ji (2005) and Gu and Ji (2005) used facial point tracking based on a Kalman filtering scheme, which is the traditional tool for solving sequential state problems. The derivation of the Kalman filter is based on a state-space model (Kalman, 1960), governed by two assumptions: (i) linearity of the model and (ii) Gaussianity of both the dynamic noise in the process equation and the measurement noise in the measurement equation. Under these assumptions, derivation of the Kalman filter leads to an algorithm that propagates the mean vector and covariance matrix of the state estimation error in an iterative manner and is optimal in the Bayesian setting. To deal with the state estimation in nonlinear dynamical systems, the extended Kalman filter was proposed, which is derived through linearization of the state-space model. However, many of the state estimation problems, including human facial expression analysis, are nonlinear and quite often non-Gaussian too. Thus, if the face undergoes a sudden or rapid movement, the prediction of features positions from Kalman filtering will be significantly off. To handle this problem, Zhang and Ji (2005) and Gu and Ji (2005) used the information about the IR-camera detected pupil location together with the output of Kalman filtering to predict facial features positions in the next frame of an input face video. To overcome these limitations of the classical Kalman filter and its extended form in general, particle filters wereproposed. For an extended overview of the various facets of particle filters see (Haykin & de Freitas, 2004).

The facial points tracking schemes employed by facial expression analyzers proposed by Pantic and colleagues (e.g., Pantic & Patras, 2006; Valstar & Pantic, 2006a) are based upon particle filtering.

The main idea behind particle filtering is to maintain a set of solutions that are an efficient representation of the conditional probability $p(a \mid Y)$, where a is the state of a temporal event to be tracked given a set of noisy observations $Y = \{y^1, \ldots, y^{-}, y\}$ up to the current time instant. This means that the distribution $p(a \mid Y)$ is represented by a set of pairs $\{(s_k, \pi_k)\}$ such that if s_k is chosen with probability equal to π_k, then it is as if s_k was drawn from $p(a \mid Y)$. By maintaining a set of solutions instead of a single estimate (as is done by Kalman filtering), particle filtering is able to track multimodal conditional probabilities $p(a \mid Y)$, and it is therefore robust to missing and inaccurate data and particularly attractive for estimation and prediction in nonlinear, non-Gaussian systems. In the particle filtering framework, our knowledge about the *a posteriori* probability $p(a \mid Y)$ is updated in a recursive way. Suppose that at a previous time instance we have a particle-based representation of the density $p(a^{-} \mid Y^{-})$, i.e., we have a collection of K particles and their corresponding weights (i.e. $\{(s_k^{-}, \pi_k^{-})\}$). Then, the classical particle filtering algorithm, so-called Condensation algorithm, can be summarized as follows (Isard & Blake, 1998).

1. Draw K particles s_k^{-} from the probability density that is represented by the collection $\{(s_k^{-}, \pi_k^{-})\}$.
2. Propagate each particle s_k^{-} with the transition probability $p(a \mid a^{-})$ in order to arrive at a collection of K particles s_k.
3. Compute the weights π_k for each particle as $\pi_k = p(y \mid s_k)$ and then normalize so that $\sum_k \pi_k = 1$.

This results in a collection of K particles and their corresponding weights $\{(s_k, \pi_k)\}$, which is an approximation of the density $p(a \mid Y)$.

The Condensation algorithm has three major drawbacks. The first one is that a large amount of particles that result from sampling from the proposal density $p(a \mid Y^{-})$ might be wasted because they are propagated into areas with small likelihood. The second problem is that the scheme ignores the fact that while a particle $s_k = \langle s_{k1}, s_{k2}, \ldots, s_{kN} \rangle$ might have low likelihood, it can easily happen that parts of it might be close to the correct solution. Finally, the third problem is that the estimation of the particle weights does not take into account the interdependences between the different parts of the state α.

Various extensions to classical Condensation algorithm have been proposed and some of those were used to track facial features. For example, Pitt and Shepard (1999) introduced Auxiliary Particle Filtering, which addresses the first drawback of the Condensation algorithm by favoring particles that end up in areas with high likelihood when propagated with the transition density $p(a \mid a^{-})$. Pantic and Patras employed this algorithm to track facial characteristic points in either face-profile- (Pantic & Patras, 2006) or in frontal-face image sequences (Pantic & Patras, 2005). To address the third drawback of the Condensation algorithm for the case of simultaneous tracking of facial components (eyes, eyebrows, nose, and mouth), Su et al. (2004) combined it with spatial belief propagation in order to enforce (pre-learned) spatial correlations between parameterizations of facial components. The extension to the Condensation algorithm used by Valstar and Pantic (2006a) for facial point tracking is the so-called Particle Filtering with Factorized Likelihoods (PFFL) proposed in (Patras & Pantic, 2004) combined with a robust color-based observation model (Patras & Pantic, 2005). This algorithm addresses the aforementioned problems inherent in the Condensation algorithm by extending the Auxiliary Particle Filtering to take into account the interdependences between the different parts of the state a. More specifically, the PFFL tracking scheme assumes that the state a can be partitioned into sub-states a_i (which, in our case, correspond to the different facial points), such that $a = \langle a_1, \ldots, a_n \rangle$. The density $p(a \mid a^{-})$, that captures the interdependencies between the locations of the facial features is estimated using a set of training data and a kernel-based density estimation scheme. As the collection of training data in question, four sets of annotated data were used containing the coordinates of facial salient points belonging to four facial components: eyebrows, eyes, nose-chin, and mouth (Patras & Pantic, 2004; Valstar & Pantic, 2006a). The underlying assumption is that correlations between the points belonging to the same facial components are more important for facial expression recognition than correlations between the points belonging to different facial components. This is consistent with psychological studies that suggest that: a) the brain processes facial expressions locally/ analytically rather than holistically whilst it identifies faces holistically (Bassili, 1978), and b) dynamic cues (expressions) are computed separately from static cues (facial proportions) (Humphreys et al., 1993). This dataset is based on 66 image sequences of 3 persons (33% female) showing 22 AUs that the facial expression analyzer proposed by Valstar and Pantic (2006a) is able to recognize. The utilized sequences are from the MMI facial expression database, part 1 (posed expressions), and they have not been used to train and test the performance of the system as a whole. Typical results of the PFFL, applied for tracking color-based templates of facial points in image sequences of faces in frontal-view are shown in Figure 5.

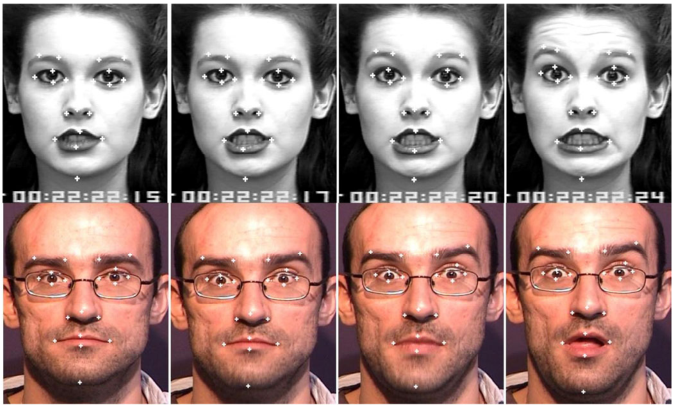

Figure 5. Example of first-effort results of the PFFL tracking scheme of Patras and Pantic (2004, 2005) for samples from the Cohn-Kanade dataset (1st row) and the MMI database (posed expressions) (2nd row)

6. Appearance-based Facial Features and Emotion Recognition

6.1 Appearance-based Facial Features

Most computer vision researchers think of motion when they consider the problem of facial expression recognition. An often cited study by Bassili (1978) shows that humans can recognize facial expressions above chance from motion, using point-light displays. However, the role of appearance-based texture information in expression recognition is like the proverbial elephant in the living room[2]. In contrast to the Bassili study in which humans were barely above chance using motion without texture, humans are nearly at ceiling for recognizing expressions from texture without motion (i.e. static photographs).

Appearance-based features include Gabor filters, integral image filters (also known as box-filters, and Haar-like filters), features based on edge-oriented histograms and those based on Active Appearance Models (Edwards et al., 1998). This set also includes spatio-temporal features like motion energy images (Essa & Pentland, 1997) and motion history images (Valstar et al., 2004), and learned image filters from independent component analysis (ICA), principal component analysis (PCA), and local feature analysis (LFA). Linear discriminant analysis (e.g., fisherfaces) is another form of learned appearance-based feature, derived from supervised learning, in contrast to the others mentioned above, which were based on unsupervised learning from the statistics of large image databases.

A common reservation about appearance-based features for expression recognition is that they are affected by lighting variation and individual differences. However, machine learning systems taking large sets of appearance-features as input, and trained on a large database of examples, are emerging as some of the most robust systems in computer vision. Machine learning combined with appearance-based features has been shown to be highly robust for tasks of face detection (Viola & Jones, 2004; Fasel et al., 2005), feature detection (Vukadinovic & Pantic, 2005; Fasel, 2006), and expression recognition (Littlewort et al., 2006). Such systems also don't suffer from issues of initialization and drift, which are major challenges for motion tracking.

The importance of appearance-based features for expression recognition is emphasized by several studies that suggest that appearance-based features may contain more information about facial expression than displacements of a set of points (Zhang et al., 1998; Donato et al., 1999), although the findings were mixed (e.g., Pantic & Patras, 2006). In any case, reducing the image to a finite set of feature displacements removes a lot of information that could be tapped for recognition. Ultimately, combining appearance-based and motion-based representations may be the most powerful, and there is some experimental evidence that this is indeed the case (e.g., Bartlett et al., 1999).

Bartlett and colleagues (Donato et al., 1999) compared a number of appearance-based representations on the task of facial action recognition using a simple nearest neighbor classifier. They found that Gabor wavelets and ICA gave better performance than PCA, LFA, Fisher's linear discriminants, and also outperformed motion flow field templates. More recent comparisons included comparisons of Gabor filters, integral image filters, and edge-oriented histograms (e.g., Whitehill & Omlin, 2006), using SVMs and AdaBoost as the classifiers. They found an interaction between feature-type and classifier, where AdaBoost performs better with integral image filters, while SVMs perform better with Gabors. The difference may be attributable to the fact that the pool of integral image filters was much

[2] Something so large that people fail to remark on it.

larger. AdaBoost performs feature selection and does well with redundancy, whereas SVMs were calculated on the full set of filters and don't do well with redundancy. Additional comparisons will be required to tease these questions apart.

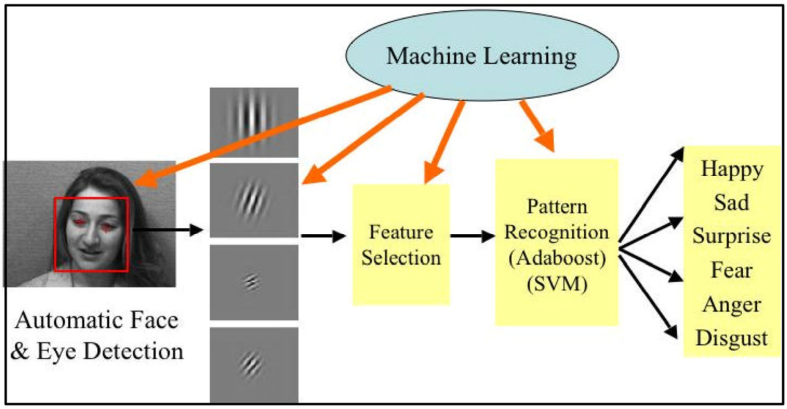
Machine Learning
Automatic Face & Eye Detection
Feature Selection
Pattern Recognition (Adaboost) (SVM)
Happy
Sad
Surprise
Fear
Anger
Disgust

Figure 6. Outline of the real-time expression recognition system of Littlewort et al. (2006)

6.2 Appearance-based Facial Affect Recognition

Here we describe the appearance-based facial expression recognition system developed by Bartlett and colleagues (Bartlett et al., 2003; Littlewort et al., 2006). The system automatically detects frontal faces in the video stream and codes each frame with respect to 7 dimensions: neutral, anger, disgust, fear, joy, sadness, surprise. The system operates in near-real-time, at about 6 frames per second on a Pentium IV. A flow diagram is shown in Figure 6. The system first performs automatic face and eye detection using the appearance-based method of Fasel et al. (2005) (see section 3). Faces are then aligned based on the automatically detected eye positions, and passed to a bank of appearance-based features. A feature selection stage extracts subsets of the features and passes them to an ensemble of classifiers which make a binary decision about each of the six basic emotions plus neutral.

Feature selection	LDA	SVM (linear)
None	44.4	88.0
PCA	80.7	75.5
Adaboost	88.2	93.3

Table 1. Comparison of feature-selection techniques in the appearance-based expression recognition system of Littlewort et al (2006). Three feature selection options are compared using LDA and SVMs as the classifier

Kernel	Adaboost	SVM	AdaSVM	LDA_{pca}
Linear	90.1	88.0	93.3	80.7
RBF		89.1	93.3	

Table 2. Comparison of classifiers in the appearance-based expression recognition system of Littlewort et al (2006). AdaSVM: Feature selection by AdaBoost followed by classification with SVM's. LDApca: Linear Discriminant analysis with feature selection based on principle component analysis, as commonly implemented in the literature

Littlewort et al. (2006) carried out empirical investigations of machine learning methods applied to this problem, including comparison of recognition engines and feature selection techniques. The feature selection techniques compared were (1) Nothing, (2) PCA, and (3) Feature selection by AdaBoost. When the output of each feature is treated as the weak classifier, AdaBoost performs feature selection, such that each new feature is the one that minimizes error, contingent on the set features that were already selected. These feature selection techniques were compared when combined with three classifiers: SVM-AdaBoost, and Linear Discriminant Analysis (LDA). The system was trained on the Cohn-Kanade dataset, and tested for generalization to new subjects using cross-validation. Results are shown in Tables 1 and 2. Best results were obtained by selecting a subset of Gabor filters using AdaBoost and then training SVMs on the outputs of the filters selected by AdaBoost. The combination of AdaBoost and SVMs enhanced both speed and accuracy of the system. The system obtained 93% accuracy on a 7-alternative forced choice. This is the highest accuracy to our knowledge on the Cohn-Kanade database, which points to the richness of appearance-based features in facial expressions. Combining this system with motion tracking and spatio-temporal analysis systems such as those developed by Pantic & Patras (2005) and Cohn et al. (2004) is a promising future direction for this research.

7. Facial Muscle Action Detection

As already mentioned in section 2.1, two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) detection and facial muscle action detection such as the AUs defined in FACS (Ekman & Friesen, 1978; Ekman et al., 2002). This section introduces recent advances in automatic facial muscle action coding.

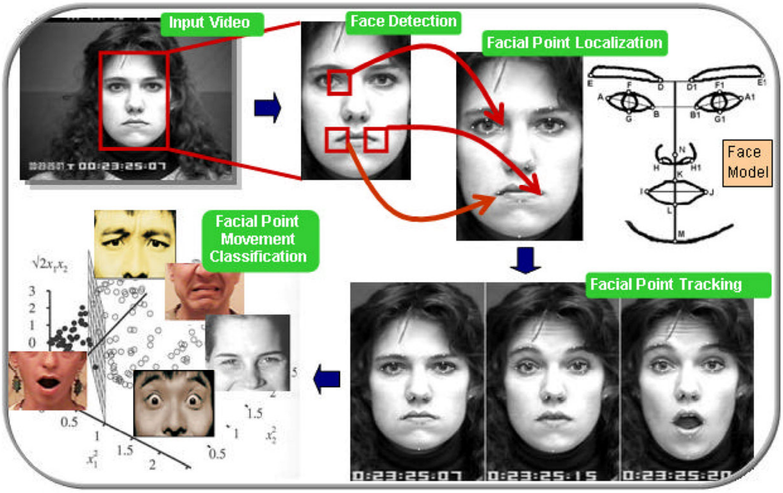

Figure 7. Outline of the AU recognition system of Valstar and Pantic (2006a)

Although FACS provides a good foundation for AU-coding of face images by human observers, achieving AU recognition by a computer is not an easy task. A problematic issue is that AUs can occur in more than 7000 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. Historically, the first attempts to encode AUs in images of faces in an automatic way were reported by Bartlett et al. (1996), Lien et al. (1998), and Pantic et al. (1998). These three research groups are still the forerunners in this research field. This section summarizes the recent work of two of those research groups, namely that of Pantic and her colleagues (section 7.1) and that of Bartlett and her colleagues (section 7.2). An application of automatic AU recognition to facial behavior analysis of pain is presented in section 7.3.

7.1 Feature-based Methods for Coding AUs and their Temporal Segments

Pantic and her colleagues reported on multiple efforts aimed at automating the analysis of facial expressions in terms of facial muscle actions that constitute the expressions. The majority of this previous work concerns geometric-feature-based methods for automatic FACS coding of face images. Early work was aimed at AU coding in static face images (Pantic & Rothkrantz, 2004) while more recent work addressed the problem of automatic AU coding in face video (Pantic & Patras, 2005, 2006; Valstar & Pantic, 2006a, 2006b). Based upon the tracked movements of facial characteristic points, as discussed in section 5, Pantic and her colleagues mainly experimented with rule-based (Pantic & Patras, 2005, 2006) and Support Vector Machine based methods (Valstar & Pantic, 2006a, 2006b) for recognition of AUs in either near frontal-view (Figure 7) or near profile-view (Figure 8) face image sequences.

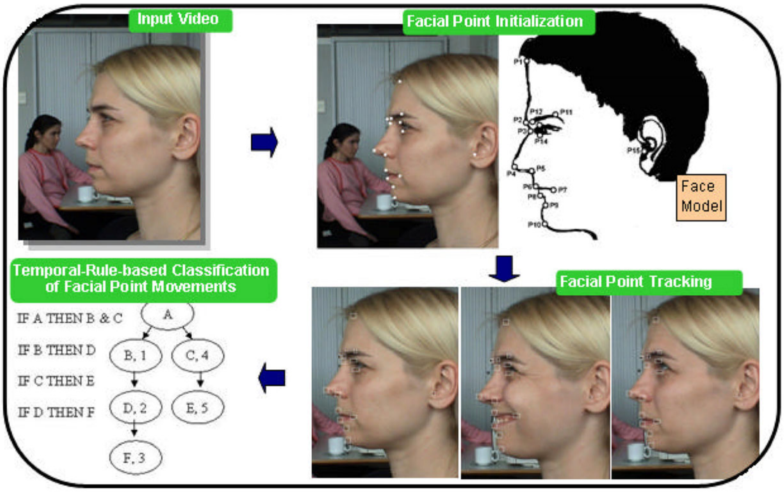

Figure 8. Outline of the AU recognition system of Pantic and Patras (2006)

As already mentioned in section 2, automatic recognition of facial expression configuration (in terms of AUs constituting the observed expression) has been the main focus of the research efforts in the field. In contrast to the methods developed elsewhere, which thus focus onto the problem of spatial modeling of facial expressions, the methods proposed by Pantic and her colleagues address the problem of temporal modeling of facial expressions as well. In other words, these methods are very suitable for encoding temporal activation patterns (onset → apex → offset) of AUs shown in an input face video. This is of importance for there is now a growing body of psychological research that argues that temporal dynamics of facial behavior (i.e., the timing and the duration of facial activity) is a critical factor for the interpretation of the observed behavior (see section 2.2). Black and Yacoob (1997) presented the earliest attempt to automatically segment prototypic facial expressions of emotions into onset, apex, and offset components. To the best of our knowledge, the only systems to date for explicit recognition of temporal segments of AUs are the ones by Pantic and colleagues (Pantic & Patras, 2005, 2006; Valstar & Pantic, 2006a, 2006b).

A basic understanding of how to achieve automatic AU detection from the profile view of the face is necessary if a technological framework for automatic AU detection from multiple views of the face is to be built. Multiple views was deemed the most promising method for achieving robust AU detection (Yacoob et al., 1998), independent of rigid head movements that can cause changes in the viewing angle and the visibility of the tracked face. To address this issue, Pantic and Patras (2006) proposed an AU recognition system from face profile-view image sequences. To the best of our knowledge this is the only such system to date.

To recognize a set of 27 AUs occurring alone or in combination in a near profile-view face image sequence, Pantic and Patras (2006) proceed under two assumptions (as defined for video samples of the MMI facial expression database, part one; Pantic et al., 2005b): (1) the input image sequence is non-occluded (left or right) near profile-view of the face with possible in-image-plane head rotations, and (2) the first frame shows a neutral expression. To make the processing robust to in-image-plane head rotations and translations as well as to small translations along the z-axis, the authors estimate a global affine transformation δ for each frame and based on it they register the current frame to the first frame of the sequence. In order to estimate the global affine transformation, they track three referential points. These are (Figure 8): the top of the forehead (P1), the tip of the nose (P4), and the ear canal entrance (P15). These points are used as the referential points because of their stability with respect to non-rigid facial movements. The global affine transformation δ is estimated as the one that minimizes the distance (in the least squares sense) between the δ-based projection of the tracked locations of the referential points and these locations in the first frame of the sequence. The rest of the facial points illustrated in Figure 8 are tracked in frames that have been compensated for the transformation δ. Changes in the position of the facial points are transformed first into a set of mid-level parameters for AU recognition. These parameters are: ***up/down(P)*** and ***inc/dec(PP')***. Parameter ***up/down(P)*** = $y(P_{t1}) - y(P_t)$, where $y(P_{t1})$ is the y-coordinate of point P in the first frame and $y(P_t)$ is the y-coordinate of point P in the current frame, describes upward and downward movements of point P. Parameter ***inc/dec(PP')*** = $PP'_{t1} - PP'_t$, where PP'_{t1} is the distance between points P and P' in the first frame and PP'_t is the distance between points P and P' in the current frame, describes the increase or decrease of the distance between points P and P'. Further, an AU can be either in:

(a) the onset phase, where the muscles are contracting and the appearance of the face changes as the facial action grows stronger, or in
(b) the apex phase, where the facial action is at its apex and there are no more changes in facial appearance due to this particular facial action, or in
(c) the offset phase, where the muscles are relaxing and the face returns to its neutral appearance, or in
(d) the neutral phase, where there are no signs of activation of this facial action.

Often the order of these phases is neutral-onset-apex-offset-neutral, but other combinations such as multiple-apex facial actions are also possible. Based on the temporal consistency of mid-level parameters, a rule-based method of Pantic and Patras encodes temporal segments (onset, apex, offset) of 27 AUs occurring alone or in combination in the input face videos. E.g., to recognize the temporal segments of AU12, the following temporal rules are used:

IF (*up/down*(P7) > ε AND *inc/dec*(P5P7) ≥ ε) THEN ***AU12-p***
IF ***AU12-p*** AND {([*up/down*(P7)]$_t$ > [*up/down*(P7)]$_{t-1}$) THEN ***AU12-onset***
IF ***AU12-p*** AND {(| [*up/down*(P7)]$_t$ – [*up/down*(P7)]$_{t-1}$ | ≤ ε) THEN ***AU12-apex***
IF ***AU12-p*** AND {([*up/down*(P7)]$_t$ < [*up/down*(P7)]$_{t-1}$) THEN ***AU12-offset***

The meaning of these rules is as follows. P7 should move upward, above its neutral-expression location, and the distance between points P5 and P7 should increase, exceeding its neutral-expression length, in order to label a frame as an "AU12 onset". In order to label a frame as "AU12 apex", the increase of the values of the relevant mid-level parameters should terminate. Once the values of these mid-level parameters begin to decrease, a frame can be labeled as "AU12 offset".

Since no other facial expression database contains images of faces in profile view, the method for AU coding in near profile-view face video was tested on MMI facial expression database only. The accuracy of the method was measured with respect to the misclassification rate of each "expressive" segment of the input sequence (Pantic & Patras, 2006). Overall, for 96 test samples, an average recognition rate of 87% was achieved sample-wise for 27 different AUs occurring alone or in combination in an input video.

For recognition of up to 22 AUs occurring alone or in combination in an input frontal-face image sequence, Valstar and Pantic (2006a) proposed a system that detects AUs and their temporal segments (neutral, onset, apex, offset) using a combination of Gentle Boost learning and Support Vector Machines (SVM). To make the processing robust to in-image-plane head rotations and translations as well as to small translations along the z-axis, the authors estimate a global affine transformation δ for each frame and based on it they register the current frame to the first frame of the sequence. To estimate δ, they track three referential points. These are: the nasal spine point (N, calculated as the midpoint between the outer corners of the nostrils H and H1, see Figure 7) and the inner corners of the eyes (B and B1, see Figure 7). The rest of the facial points illustrated in Figure 7 are tracked in frames that have been compensated for the transformation δ. Typical tracking results are shown in Figure 5. Then, for all characteristic facial points *Pi* depicted in Figure 7, where *i* = [1 : 20], they compute two the displacement of *Pi* in y- and x-direction for every frame *t*. Then, for all pairs of points *Pi* and *Pj*, where *i* ≠ *j* and *i,j* = [1 : 20], they compute in each frame the distances between the points and the increase/decrease of the distances in correspondence to the first frame. Finally, they compute the first time derivative *df/dt* of all features defined above, resulting in a set of 1220 features per frame.

They use further Gentle Boost (Friedman et al., 2000) to select the most informative features for every class *c* ∈ *C*, where *C* = {AU1, AU2, AU4, AU5, AU6, AU7, AU43, AU45, AU46, AU9, AU10, AU12, AU13, AU15, AU16, AU18, AU20, AU22, AU24, AU25, AU26, AU27}. An advantage of feature selection by Gentle Boost is that features are selected depending on the features that have been already selected. In feature selection by Gentle Boost, each feature is treated as a weak classifier. Gentle Boost selects the best of those classifiers and then boosts the weights using the training examples to weight the errors more. The next feature is selected as the one that gives the best performance on the errors of the previously selected features. At each step, it can be shown that the chosen feature is uncorrelated with the output of the previously selected features. As shown by Littlewort et al. (2006), when SVMs are trained using the features selected by a boosting algorithm, they perform better.

To detect 22 AUs occurring alone or in combination in the current frame of the input sequence (i.e., to classify the current frame into one or more of the $c \in C$), Valstar and Pantic use 22 separate SVMs to perform binary decision tasks using one-versus-all partitioning of data resulting from the feature selection stage. More specifically, they use the most informative features selected by Gentle Boost for the relevant AU (i.e., the relevant $c \in C$) to train and test the binary SVM classifier specialized in recognition of that AU. They use radial basis function (RBF) kernel employing a unit-width Gaussian. This choice has been influenced by research findings of Bartlett et al. (2006) and Littlewort et al. (2006), who provided experimental evidence that Gaussian RBF kernels are very well suited for AU detection, especially when the SVM-based classification is preceded by an ensemble-learning-based feature selection.

As every facial action can be divided into four temporal segments (neutral, onset, apex, offset), Valstar and Pantic consider the problem to be a four-valued multi-class classification problem. They use a one-versus-one approach to multi-class SVMs (mc-SVMs). In this approach, for each AU and every pair of temporal segments, a separate sub-classifier specialized in the discrimination between the two temporal segments is trained. This results in $\sum_i i = 6$ sub-classifiers that need to be trained ($i = [1 : C - 1]$, $C = \{$neutral, onset, apex, offset$\}$). For each frame t of an input image sequence, every sub-classifier returns a prediction of the class $c \in C$, and a majority vote is cast to determine the final output c_t of the mc-SVM for the current frame t. To train the sub-classifiers, Valstar and Pantic apply the following procedure using the same set of features that was used for AU detection (see equations (1)–(5) above). For each classifier separating classes $c_i, c_j \in C$ they apply Gentle Boost, resulting in a set of selected features $G_{i,j}$. They use $G_{i,j}$ to train the sub-classifier specialized in discriminating between the two temporal segments in question ($c_i, c_j \in C$).

The system achieved average recognition rates of 91% and 97% for samples from the Cohn-Kanade facial expression database (Kanade et al., 2000) and, respectively, the MMI facial expression database (Pantic et al. 2005b), 84% when trained on the MMI and tested on the Cohn-Kanade database samples, and 52% when trained on the MMI database samples and tested on the spontaneous-data-part of the MMI database.

Experiments concerning recognition of facial expression temporal activation patterns (onset → apex → offset) were conducted on the MMI database only, since the sequences in the Cohn-Kanade database end at the apex. On average, 95% of temporal patterns of AU activation were detected correctly by their system. The system successfully detected the duration of most AUs as well, with a shift of less than 2 frames in average. However, for AU6 and AU7, the measurement of the duration of the activation was over 60% off from the actual duration. It seems that human observers detect activation of these AUs not only based on the presence of a certain movement (like an upward movement of the lower eyelid), but also based on the appearance of the facial region around the eye corner (like the crow feet wrinkles in the case of AU6). Such an appearance change may be of a different duration from the movement of the eyelid, resulting in an erroneous estimation of AU duration by the system that takes only facial movements into account. As mentioned above, using both geometric and appearance features might be the best choice in the case of such AUs.

7.2 Appearance-based Methods for AU Coding

Here we describe an appearance-based system for fully automated facial action coding developed by Bartlett and colleagues (Bartlett et al. 2003, 2006), and show preliminary results when applied to spontaneous expressions. This extends a line of research developed in collaboration with Paul Ekman and Terry Sejnowski (e.g., Bartlett et al., 1996, 1999). The system is the same as the one described in Section 6.1, with the exception that the system was trained to detect facial actions instead of basic emotions. An overview is shown in Figure 9. It is user independent and operates in near-real time, at about 6 frames per second on a Pentium IV. The system detects 30 AUs, and performance measures are available for 20 of them, below. Bartlett and colleagues (2006) also found that this system captures information about AU intensity that can be employed for analyzing facial expression dynamics.

Appearance-based approaches to AU recognition such as the one presented here, differ from those of Pantic (e.g., Pantic & Rothkrantz, 2004a) and Cohn (e.g., Tian et al., 2001), in that instead of employing heuristic, rule-based methods, and/or designing special purpose detectors for each AU, these methods employ machine learning in a general purpose system that can detect any AU given a set of labeled training data. Hence the limiting factor in appearance-based machine learning approaches is having enough labeled examples for a robust system. Previous explorations of this idea showed that, given accurate 3D alignment, at least 50 examples are needed for moderate performance (in the 80% range), and over 200 examples are needed to achieve high precision (Bartlett et al., 2003). Another prototype appearance-based system for fully automated AU coding was presented by Kapoor et al. (2003). This system used infrared eye tracking to register face images. The recognition component is similar to the one presented here, employing machine learning techniques on feature-based representations, where Kapoor et al. used PCA (eigenfeatures) as the feature vector to be passed to an SVM. As mentioned in Section 6.1, we previously found that PCA was a much less effective representation than Gabor wavelets for facial action recognition with SVMs. An appearance-based system was also developed by Tong et al. (2006). They applied a dynamic Bayesian model to the output of a front-end AU recognition system based on the one developed in the Bartlett's laboratory. While Tong et al. showed that AU recognition benefits from learning causal relations between AUs in the training database, the analysis was developed and tested on a posed expression database. It will be important to extend such work to spontaneous expressions for the reasons described in Section 2.2.

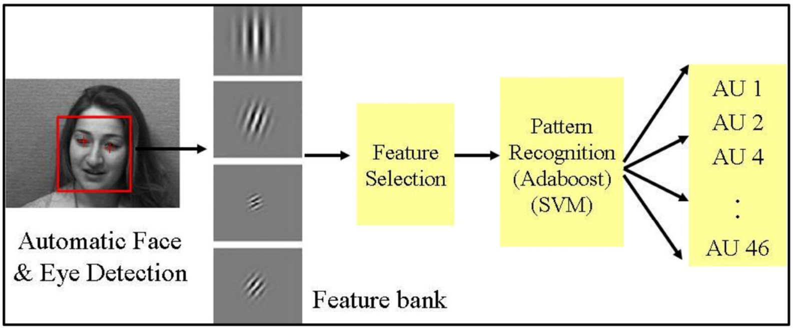

Figure 9. Outline of the Appearance-based facial action detection system of Bartlett et al. (2006)

Here we show performance of the system of Bartlett et al. (2006) for recognizing facial actions in posed and spontaneous expressions (Figure 10). The system was trained on both the Cohn-Kanade and Ekman-Hager datasets. The combined dataset contained 2568 training examples from 119 subjects. Performance presented here was for training and testing on 20 AUs. Separate binary classifiers, one for each AU, were trained to detect the presence of the AU regardless of the co-occurring AUs. Positive examples consisted of the last frame of each sequence which contained the expression apex. Negative examples consisted of all apex frames that did not contain the target AU plus neutral images obtained from the first frame of each sequence, for a total of 2568-N negative examples for each AU.

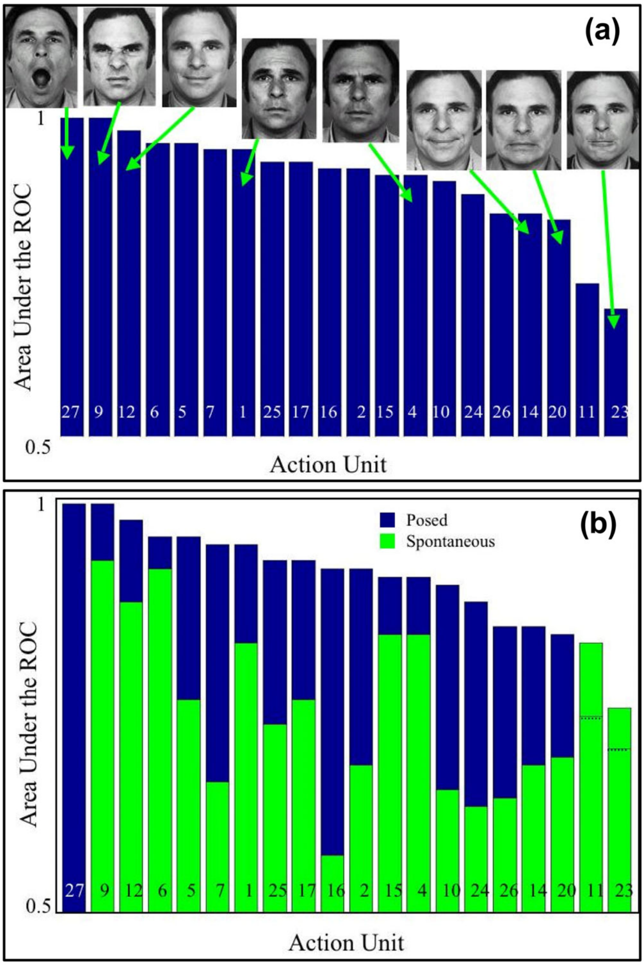

Figure 10. System performance (area under the ROC) for the AU detection system of Bartlett et al. (2006): (a) posed facial actions (sorted in order of detection performance), and (b) spontaneous facial actions (performance is overlayed on the posed results of (a); there were no spontaneous examples of AU 27 in this sample)

We first report performance for generalization to novel subjects *within* the Cohn-Kanade and Ekman-Hager databases. Generalization to new subjects was tested using leave-one-subject-out cross-validation in which all images of the test subject were excluded from training. The system obtained a mean of 91% agreement with human FACS labels. Overall percent correct can be an unreliable measure of performance, however, since it depends on the proportion of targets to non-targets, and also on the decision threshold. In this test, there was a far greater number of non-targets than targets, since targets were images containing the desired AU (N), and non-targets were all images not containing the desired AU (2568-N). A more reliable performance measure is area under the ROC (receiver-operator characteristic curve, or A'). This curve is obtained by plotting hit rate (true positives) against false alarm rate (false positives) as the decision threshold varies. A' is equivalent to percent correct in a 2-alternative forced choice task, in which the system must choose which of two options contains the target on each trial. Mean A' for the posed expressions was 92.6.

A correlation analysis was performed in order to explicitly measure the relationship between the output margin and expression intensity. Ground truth for AU intensity was measured as follows: Five certified FACS coders labeled the action intensity for 108 images from the Ekman-Hager database, using the A-E scale of the FACS coding manual, where A is lowest, and E is highest. The images were four upper-face actions (1, 2, 4, 5) and two lower-face actions (10, 20), displayed by 6 subjects. We first measured the degree to which expert FACS coders agree with each other on intensity. Correlations were computed between intensity scores by each pair of experts, and the mean correlation was computed across all expert pairs. Correlations were computed separately for each display subject and each AU, and then means were computed across display subjects. Mean correlation between expert FACS coders within subject was 0.84.

	Action unit						
	1	2	4	5	10	20	Mean
Expert-Expert	.92	.77	.85	.72	.88	.88	.84
SVM-Expert	.90	.80	.84	.86	.79	.79	.83

Table 3. Correlation of SVM margin with intensity codes from human FACS experts

Correlations of the automated system with the human expert intensity scores were next computed. The SVMs were retrained on the even-numbered subjects of the Cohn-Kanade and Ekman-Hager datasets, and then tested on the odd-numbered subjects of the Ekman-Hager set, and vice versa. Correlations were computed between the SVM margin and the intensity ratings of each of the five expert coders. The results are shown in Table 3. Overall mean correlation between the SVM margin and the expert FACS coders was 0.83, which was nearly as high as the human-human correlation of .84. Similar findings were obtained using an AdaBoost classifier, where the AdaBoost output, which is the likelihood ratio of target/nontarget, correlated positively with human FACS intensity scores (Bartlett et al., 2004).

The system therefore is able to provide information about facial expression dynamics in terms of the frame-by-frame intensity information. This information can be exploited for deciding the presence of an AU and decoding the onset, apex, and offset. It will also enable studying the dynamics of facial behavior. As explained in section 2, enabling investigations into the dynamics of facial expression would allow researchers to directly address a number of questions key to understanding the nature of the human emotional and expressive systems, and their roles interpersonal interaction, development, and psychopathology.

We next tested the system on the RU-FACS Dataset of spontaneous expressions described in section 2.5. The results are shown in Figure 10. The dataset included speech related mouth and face movements, and significant amounts of in-plane and in-depth rotations. Yaw, pitch, and roll ranged from -30 to 20 degrees. Preliminary recognition results are presented for 12 subjects. This data contained a total of 1689 labeled events, consisting of 33 distinct action units, 19 of which were AUs for which we had trained classifiers. All detected faces were passed to the AU recognition system. Faces were detected in 95% of the video frames. Most non-detects occurred when there was head rotations beyond ±10^0 or partial occlusion.

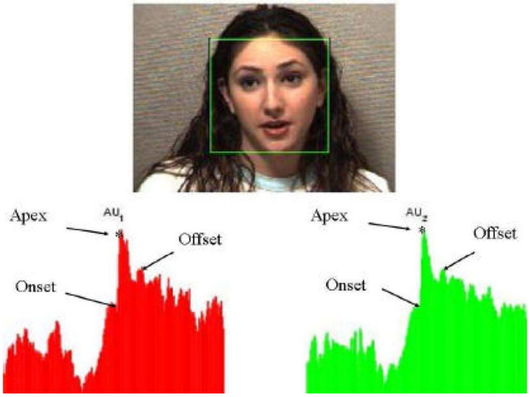

Figure 11. Output of automatic FACS coding system from Bartlett et al. (2006). Frame-by-frame outputs are shown for AU 1 and AU 2 (brow raise) for 200 frames of video. The output is the distance to the separating hyperplane of the SVM. Human codes (onset, apex, and offset frame) are overlayed for comparison

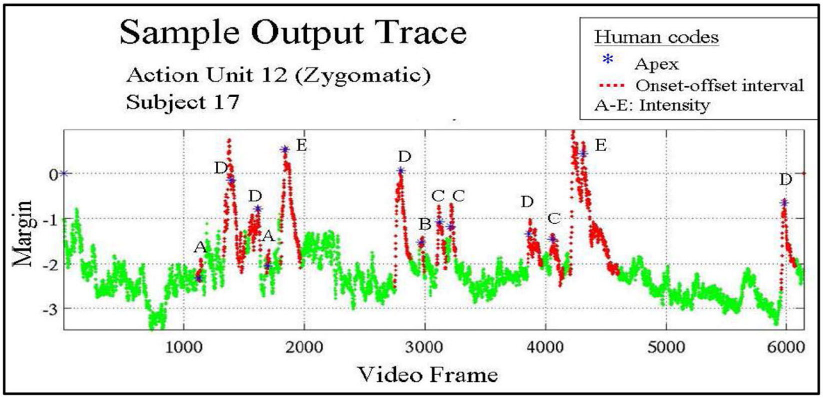

Figure 12. Output trajectory for a 2′ 20′′ video (6000 frames), for one subject and one action unit. Shown is the margin (the distance to the separating hyperplane). The human FACS labels are overlaid for comparison: Frames within the onset and offset of the AU are shown in red. Stars indicate the AU apex frame. Letters A-E indicate AU intensity, with E highest

Example system outputs are shown in Figure 11 and 12. The system obtained a mean of 93% correct detection rate across the 19 AUs in the spontaneous expression data. As explained above, however, percent correct can be misleading when there are unequal numbers of targets and nontargets. Mean area under the ROC for the spontaneous action units was .75 (and thus percent correct on a 2-alternative forced choice would be 75%). This figure is nevertheless encouraging, as it shows that fully automated systems can indeed get a signal about facial actions, despite generalizing from posed to spontaneous examples, and despite the presence of noise from speech and out-of-plane head rotations. As with the posed expression data, the SVM margin correlated positively with AU intensity in the spontaneous data (Figure 12). Mean correlation of AU 12 with FACS intensity score was .75, and the mean over eight AUs tested was 0.35.

7.3 Automatic Detection of Pain

The automated AU recognition system described above was applied to spontaneous facial expressions of pain (Littlewort et al., 2006b). The task was to differentiate faked from real pain expressions using the automated AU detector. Human subjects were videotaped while they submerged their hand in a bath of water for three minutes. Each subject experienced three experimental conditions: baseline, real pain, and posed pain. In the real pain condition, the water was 3 degrees Celsius, whereas in the baseline and posed pain conditions the water was 20 degrees Celsius. The video was coded for AUs by both human and computer. Our initial goal was to correctly determine which experimental condition is shown in a 60 second clip from a previously unseen subject. For this study, we trained individual AU classifiers on 3000 single frames selected from three datasets: two posed expression sets, the Cohn-Kanade and the Ekman-Hager datasets, and the RU-FACS dataset of spontaneous expression data. We trained linear SVM for each of 20 AUs, in one versus all mode, irrespective of combinations with other AUs. The output of the system was a real valued number indicating the distance to the separating hyperplane for each classifier. Applying this system to the pain video data produced a 20 channel output stream, consisting of one real value for each learned AU, for each frame of the video. This data was further analyzed to predict the difference between expressions of real pain and fake pain. The 20-channel output streams were passed to another set of three SVMs, trained to detect real pain, fake pain, and baseline. In a preliminary analysis of 5 subjects tested with cross-validation, the system correctly identified the experimental condition (posed pain, real pain, and baseline) for 93% of samples in a 3-way forced choice. The 2-way performance for fake versus real pain was 90%. This is considerably higher than the performance of naive human observers, who are near chance for identifying faked pain (Hadjistavropoulos et al., 1996).

8. Challenges, Opportunities and Recommendations

Automating the analysis of facial signals, especially rapid facial signals (facial expressions) is important to realize more natural, context-sensitive (e.g., affective) human-computer interaction, to advance studies on human emotion and affective computing, and to boost numerous applications in fields as diverse as security, medicine, and education. This chapter introduced recent advances in machine analysis of facial expressions and summarized the recent work of two forerunning research groups in this research field, namely that of Pantic and her colleagues and that of Bartlett and her colleagues.

In summary, although most of the facial expression analyzers developed so far target human facial affect analysis and attempt to recognize a small set of prototypic emotional facial expressions like happiness and anger (Pantic et al., 2005a), some progress has been made in addressing a number of other scientific challenges that are considered essential for realization of machine understanding of human facial behavior. First of all, the research on automatic detection of facial muscle actions, which produce facial expressions, witnessed a significant progress in the past years. A number of promising prototype systems have been proposed recently that can recognize up to 27 AUs (from a total of 44 AUs) in either (near-) frontal view or profile view face image sequences (section 7 of this chapter; Tian et al. 2005). Further, although the vast majority of the past work in the field does not make an effort to explicitly analyze the properties of facial expression temporal dynamics, a few approaches to automatic segmentation of AU activation into temporal segments (neutral, onset, apex, offset) have been recently proposed (section 7 of this chapter). Also, even though most of the past work on automatic facial expression analysis is aimed at the analysis of posed (deliberately displayed) facial expressions, a few efforts were recently reported on machine analysis of spontaneous facial expressions (section 7 of this chapter; Cohn et al., 2004; Valstar et al., 2006; Bartlett et al., 2006). In addition, exceptions from the overall state of the art in the field include a few works towards detection of attitudinal and non-basic affective states such as attentiveness, fatigue, and pain (section 7 of this chapter; El Kaliouby & Robinson, 2004; Gu & Ji, 2004), a few works on context-sensitive (e.g., user-profiled) interpretation of facial expressions (Fasel et al., 2004; Pantic & Rothkrantz, 2004b), and an attempt to explicitly discern in an automatic way spontaneous from volitionally displayed facial behavior (Valstar et al., 2006). However, many research questions raised in section 2 of this chapter remain unanswered and a lot of research has yet to be done.

When it comes to automatic AU detection, existing methods do not yet recognize the full range of facial behavior (i.e. all 44 AUs defined in FACS). For machine learning approaches, increasing the number of detected AUs boils down to obtaining labeled training data. To date, Bartlett's team has means to detect 30 AUs, and do not yet have sufficient labeled data for the other AUs. In general, examples from over 50 subjects are needed. Regarding feature tracking approaches, a way to deal with this problem is to look at diverse facial features. Although it has been reported that methods based on geometric features are usually outperformed by those based on appearance features, recent studies like that of Pantic & Patras (2006), Valstar and Pantic (2006a), and those presented in this chapter, show that this claim does not always hold. We believe, however, that further research efforts toward combining both approaches are necessary if the full range of human facial behavior is to be coded in an automatic way.

Existing methods for machine analysis of facial expressions discussed throughout this chapter assume that the input data are near frontal- or profile-view face image sequences showing facial displays that always begin with a neutral state. In reality, such assumption cannot be made. The discussed facial expression analyzers were tested on spontaneously occurring facial behavior, and do indeed extract information about facial behavior in less constrained conditions such as an interview setting (e.g., Bartlett et al., 2006; Valstar et al, 2006). However deployment of existing methods in fully unconstrained environments is still in the relatively distant future. Development of robust face detectors, head-, and facial feature trackers, which will be robust to variations in both face orientation relative to the camera, occlusions, and scene complexity like the presence of other people and dynamic background, forms the first step in the realization of facial expression analyzers capable of handling unconstrained environments.

Consequently, if we consider the state of the art in face detection and facial feature localization and tracking, noisy and partial data should be expected. As remarked by Pantic and Rothkrantz (2003), a facial expression analyzer should be able to deal with these imperfect data and to generate its conclusion so that the certainty associated with it varies with the certainty of face and facial point localization and tracking data. For example, the PFFL point tracker proposed by Patras and Pantic (2004, 2005) is very robust to noise, occlusion, clutter and changes in lighting conditions and it deals with inaccuracies in facial point tracking using a memory-based process that takes into account the dynamics of facial expressions. Nonetheless, this tracking scheme is not 100% accurate. Yet, the method proposed by Valstar and Pantic (2006a), which utilizes the PFFL point tracker, does not calculate the output data certainty by propagating the input data certainty (i.e., the certainty of facial point tracking). The only work in the field that addresses this issue is that of Pantic and Rothkrantz (2004a). It investigates AU recognition from static face images and explores the use of measures that can express the confidence in facial point localization and that can facilitate assessment of the certainty of the performed AU recognition. Another way of generating facial-expression-analysis output such that the certainty associated with it varies in accordance to the input data is to consider the time-instance versus time-scale dimension of facial behavior (Pantic & Rothkrantz, 2003). By considering previously observed data (time scale) with respect to the current data (time instance), a statistical prediction and its probability might be derived about both the information that may have been lost due to malfunctioning / inaccuracy of the camera (or a part of facial expression analyzer) and the currently displayed facial expression. Probabilistic graphical models, like Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN) are well suited for accomplishing this (Pantic et al., 2005a). These models can handle noisy features, temporal information, and partial data by probabilistic inference.

It remains unresolved, however, how the grammar of facial behavior can be learned (in a human-centered manner or in an activity-centered manner) and how this information can be properly represented and used to handle ambiguities in the observation data (Pantic et al., 2005a). Another related issue that should be addressed is how to include information about the context (environment, user, user's task) in which the observed expressive behavior was displayed so that a context-sensitive analysis of facial behavior can be achieved. These aspects of machine analysis of facial expressions form the main focus of the current and future research in the field. Yet, since the complexity of these issues concerned with the interpretation of human behavior at a deeper level is tremendous and spans several different disciplines in computer and social sciences, we believe that a large, focused, interdisciplinary, international program directed towards computer understanding of human behavioral patterns (as shown by means of facial expressions and other modes of social interaction) should be established if we are to experience true breakthroughs in this and the related research fields.

9. References

Ambadar, Z., Schooler, J. & Cohn, J.F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, Vol. 16, No. 5, pp. 403-410.

Ambady, N. & Rosenthal, R. (1992). Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin*, Vol. 111, No. 2, pp. 256-274.

Anderson, K. & McOwan, P.W. (2006). A Real-Time Automated System for Recognition of Human Facial Expressions. *IEEE Trans. Systems, Man, and Cybernetics – Part B*, Vol. 36, No. 1, pp. 96-105.

Barron, J., Fleet, D. & Beauchemin, S. (1994). Performance of optical flow techniques. *J. Computer Vision*, Vol. 12, No. 1, pp. 43-78.

Bartlett, M.S., Hager, J. C., Ekman, P. & Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, Vol. 36, No. 2, pp. 253-263.

Bartlett, M.S., Littlewort, G., Braathen, B., Sejnowski, T.J., & Movellan, J.R. (2003a). A prototype for automatic recognition of spontaneous facial actions. *Advances in Neural Information Processing Systems*, Vol. 15, pp. 1271-1278.

Bartlett, M.S., Littlewort, G., Fasel, I. & Movellan, J.R. (2003b). Real time face detection and expression recognition: Development and application to human-computer interaction, *Proc. CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, p. 6.

Bartlett, M.S., Littlewort, G., Frank, M.G., Lainscsek, C., Fasel, I. & Movellan, J. (2005). Recognizing facial expression: machine learning and application to spontaneous behavior, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 568-573.

Bartlett, M.S., Littlewort, G., Frank, M.G., Lainscsek, C., Fasel, I. & Movellan, J. (2006). Fully automatic facial action recognition in spontaneous behavior, *Proc. IEEE Conf. Automatic Face & Gesture Recognition*, pp. 223-230.

Bartlett, M., Littlewort, G., Lainscsek, C., Fasel, I. & Movellan, J. (2004). Machine learning methods for fully automatic recognition of facial expressions and facial actions, *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, Vol. 1, pp. 592-597.

Bartlett, M.S., Viola, P.A., Sejnowski, T.J., Golomb, B.A., Larsen, J., Hager, J.C. & Ekman, P. (1996). Classifying facial actions, *Advances in Neural Information Processing Systems 8*, pp. 823-829.

Bassili, J.N. (1978). Facial motion in the perception of faces and of emotional expression. *J. Experimental Psychology*, Vol. 4, No. 3, pp. 373-379.

Black, M. & Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *Computer Vision*, Vol. 25, No. 1, pp. 23-48.

Bowyer, K.W. (2004). Face Recognition Technology – Security vs. Privacy. *IEEE Technology and Society Magazine*, Vol. 23, No. 1, pp. 9-19.

Brodal, A. (1981). *Neurological anatomy: In relation to clinical medicine*. Oxford University Press, New York, USA.

Chang, Y., Hu, C., Feris, R. & Turk, M. (2006). Manifold based analysis of facial expression. *J. Image & Vision Computing,* Vol. 24, No. 6, pp. 605-614.

Chen, L., Zhang, L., Zhang, H. & Abdel-Mottaleb, M. (2004). 3D Shape Constraint for Facial Feature Localization using Probabilistic-like Output, *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures*, pp. 302-307.

Cohen, I., Sebe, N., Garg, A., Chen, L.S. & Huang, T.S. (2003). Facial expression recognition from video sequences – temporal and static modelling. *Computer Vision and Image Understanding*, Vol. 91, pp. 160-187.

Cohen, M.M. (2006). *Perspectives on the Face*, Oxford University Press, Oxford, UK:

Cohn, J.F. (2006). Foundations of human computing: Facial expression and emotion, *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 233-238.

Cohn, J.F. & Ekman, P. (2005). Measuring facial actions, In: *The New Handbook of Methods in Nonverbal Behavior Research*, Harrigan, J.A., Rosenthal, R. & Scherer, K., (Eds.), pp. 9-64, Oxford University Press, New York, USA.

Cohn, J.F., Reed, L.I., Ambadar, Z., Xiao, J. and Moriyama, T. (2004). Automatic analysis and recognition of brow actions in spontaneous facial behavior, *Proc. IEEE Int'l Conf. Systems, Man & Cybernetics*, pp. 610-616.

Cohn, J.F. & Schmidt, K.L. (2004). The timing of facial motion in posed and spontaneous smiles. *J. Wavelets, Multi-resolution & Information Processing*, Vol. 2, No. 2, pp. 121-132.

Costa, M., Dinsbach, W., Manstead, A.S.R. & Bitti P.E.R. (2001). Social presence, embarrassment, and nonverbal behaviour. *J. Bonverbal Behaviour*, Vol. 25., No. 4, pp. 225-240.

Craig, K., Hyde, S.A. & Patrick, C.J. (1991). Genuine, suppressed, and faked facial behavior during exacerbation of chronic low back pain. *Pain*, Vol. 46, pp. 161–172.

Cunningham, D.W., Kleiner, M., Wallraven, C. & Bülthoff, H.H. (2004). The components of conversational facial expressions, *Proc. ACM Int'l Symposium on Applied Perception in Graphics and Visualization*, pp. 143-149.

Darwin, C. (1872/1998). *The expression of the emotions in man and animals*, Oxford University Press, New York, USA.

Daugman, J. (1988). Complete discrete 2D Gabor transform by neural networks for image analysis and compression. *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. 36, pp. 1169-1179.

DeCarlo, D. & Metaxas, D. (1996). The integration of optical flow and deformable models with applications to human face shape and motion estimation, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 231-238.

Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P. & Sejnowski, T.J. (1999). Classifying facial actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10, pp. 974-989.

Duchenne de Bologne, G.B. (1862/1990). *Mechanisme de la Physionomie Humaine*, Jules Renouard Libraire, Paris, France, 1862. (Translation: *The Mechanism of Human Facial Expression*, Cambridge University Press, New York, USA, 1990).

Edwards, G.J., Cootes, T.F. & Taylor, C.J. (1998). Face Recognition Using Active Appearance Models, *Proc. European Conf. Computer Vision*, Vol. 2, pp. 581-695.

Ekman, P. (1991). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W.W. Norton, New York, USA.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals New York Academy of sciences,* Vol. 1000, pp. 205-221.

Ekman, P. (1989). The argument and evidence about universals in facial expressions of emotion. In: *Psychological methods in criminal investigation and evidence*, Raskin, D.C., (Ed.), pp. 297–332, Springer, New York, USA.

Ekman, P. & Friesen, W.V. (1969). The repertoire of nonverbal behavior. *Semiotica*, Vol. 1, pp. 49-98.

Ekman, P. & Friesen, W.V. (1975). *Unmasking the face*, Prentice-Hall, New Jersey, USA.

Ekman, P. & Friesen, W.V. (1978). *Facial Action Coding System*, Consulting Psychologist Press, Palo Alto, USA.

Ekman, P., Friesen, W.V. & Hager, J.C. (2002). *Facial Action Coding System*, A Human Face, Salt Lake City, USA.

Ekman, P., Hager, J.C., Methvin, C.H. & Irwin, W. (1999). Ekman-Hager Facial Action Exemplars (unpublished), Human Interaction Lab, University of California, San Francisco, USA.

Ekman, P., Huang, T.S., Sejnowski, T.J. & Hager, J.C., (Eds.), (1993). *NSF Understanding the Face*, A Human Face eStore, Salt Lake City, USA, (see Library).

Ekman, P. & Rosenberg, E.L., (Eds.), (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the FACS*, Oxford University Press, Oxford, UK.

El Kaliouby, R. & Robinson, P. (2004). Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* Vol. 3, p. 154.

Essa, I. & Pentland, A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 757-763.

Fasel, B., Monay, F. & Gatica-Perez, D. (2004). Latent semantic analysis of facial action codes for automatic facial expression recognition, *Proc. ACM Int'l Workshop on Multimedia Information Retrieval*, pp. 181-188.

Fasel, I.R. (2006). *Learning Real-Time Object Detectors: Probabilistic Generative Approaches*. PhD thesis, Department of Cognitive Science, University of California, San Diego, USA.

Fasel, I.R., Fortenberry, B. & Movellan, J.R. (2005). A generative framework for real time object detection and classification. *Int'l J Computer Vision and Image Understanding*, Vol. 98, No. 1, pp. 181-210.

Frank, M.G. & Ekman, P. (2004). Appearing truthful generalizes across different deception situations. *Journal of personality and social psychology*, Vol. 86, pp. 486–495.

Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, Vol. 28, No 2, pp. 337-374.

Gokturk, S.B., Bouguet, J.Y., Tomasi, C. & Girod, B. (2002). Model-based face tracking for view independent facial expression recognition, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 272-278.

Gross, R., Matthews, I. & Baker, S. (2006). Active appearance models with occlusion. *J. Image & Vision Computing,* Vol. 24, No. 6, pp. 593-604.

Gu, H. & Ji, Q. (2004). An automated face reader for fatigue detection, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 111-116.

Gu, H. & Ji, Q. (2005). Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications*, Vol. 16, No. 2, pp. 105-115.

Guo, G. & Dyer, C.R. (2005). Learning From Examples in the Small Sample Case – Face Expression Recognition. *IEEE Trans. Systems, Man, and Cybernetics – Part B*, Vol. 35, No. 3, pp. 477-488.

Hadjistavropoulos, H.D., Craig, K.D., Hadjistavropoulos, T. & Poole, G.D. (1996). Subjective judgments of deception in pain expression: Accuracy and errors. *Pain,* Vol. 65, pp. 247-254.

Haykin, S. & de Freitas, N., (Eds.), (2004). *Special Issue on Sequential State Estimation. Proceedings of the IEEE*, vol. 92, No. 3, pp. 399-574.

Heller, M. & Haynal, V. (1997). Depression and suicide faces. In: *What the Face Reveals*, Ekman, P. & Rosenberg, E., (Eds.), pp. 339-407, Oxford University Press, New York, USA.

Hess, U., Blairy, S. & Kleck, R.E. (1997). The intensity of emotional facial expressions and decoding accuracy. *J. Nonverbal Behaviour*, Vol. 21, No. 4, pp. 241-257.

Holden, E. & Owens, R. (2002). Automatic Facial Point Detection, *Proc. Asian Conf. Computer Vision*, vol. 2, pp 731-736.

Huang, K.S. & Trivedi, M.M. (2004). Robust Real-Time Detection, Tracking, and Pose Estimation of Faces in Video Streams, *Proc. IEEE Int'l Conf. Pattern Recognition*, Vol. 3, pp. 965-968.

Humphreys, G.W., Donnelly, N. & Riddoch, M.J. (1993). Expression is computed separately from facial identity and it is computed separately for moving and static faces – Neuropsychological evidence. *Neuropsychologia*, Vol. 31, pp. 173-181.

Isard, M. & Blake, A. (1998). Condensation - conditional density propagation for visual tracking. *J. Computer Vision*, Vol. 29, No. 1, pp. 5-28.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.*, Vol. 82, pp. 35-45.

Kanade, T., Cohn, J.F. & Tian, Y. (2000). Comprehensive database for facial expression analysis, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 46-53.

Kapoor, A., Qi Y. & Picard, R.W. (2003). Fully automatic upper facial action recognition, Proc. *IEEE Int'l Workshop on Analysis and Modeling of Faces and Gestures*, pp. 195-202.

Keltner, D. (1997). Signs of Appeasement: Evidence for distinct displays of embarrassment, amusement, and shame, In: *What the Face Reveals*, Ekman, P. & Rosenberg, E., (Eds.), pp. 133-160, Oxford University Press, New York, USA.

Keltner, D. & Ekman, P. (2000). Facial Expression of Emotion, In: *Handbook of Emotions*, Lewis, M. & Haviland-Jones, J.M., (Eds.), pp. 236-249, Guilford Press, New York, USA.

Kimura, S. & Yachida, M. (1997). Facial expression recognition and its degree estimation, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 295-300.

Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R.P. & Konen, W. (1992). Distortion Invariant Object Recognition in the Dynamik Link Architecture. *IEEE Transactions on Computers*, Vol. 42, No. 3, pp. 300-311.

Li, S.Z. & Jain, A.K., (Eds.), (2005). *Handbook of Face Recognition*, Springer, New York, USA.

Lien, J.J.J., Kanade, T., Cohn, J.F. & Li, C.C. (1998). Subtly different facial expression recognition and expression intensity estimation, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 853-859.

Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J. & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *J. Image & Vision Computing,* Vol. 24, No. 6, pp. 615-625.

Littlewort, G., Bartlett, M.S. & Lee, K. (2006b). Faces of Pain: Automated measurement of spontaneous facial expressions of genuine and posed pain, *Proc. 13th Joint Symposium on Neural Computation*, p. 1.

Lucas, B.D. & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision, *Proc. Conf. Artificial Intelligence*, pp. 674-679.

Lyons, M.J.; Budynek, J. &Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, No. 12, pp. 1357-1362.

Maat, L. & Pantic, M. (2006). Gaze-X: Adaptive affective multimodal interface for single-user office scenarios, *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 171-178.

Martinez, A. M. (2003). Matching expression variant faces. *Vision Research*, Vol. 43, No. 9, pp. 1047-1060.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions*, Vol. E74, No. 10, pp. 3474-3483.

Meihlke, A. (1973). *Surgery of the facial nerve*. Saunders, Philadelphia, USA.

Moghaddam, B. & Pentland, A. (1997). Probabilistic Visual Learning for Object Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 696-710.

Nock, H.J., Iyengar, G. & Neti, C. (2004). Multimodal processing by finding common cause. *Communications of the ACM*, Vol. 47, No. 1, pp. 51-56.

O'Toole, A.J., Harms, J., Snow, S.L., Hurst, D.R., Pappas, M.R., Ayyad, J.H. & Abdi, H. (2005). A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine*, Vol. 27, No. 5, pp. 812–816.

Osadchy, M., Jacobs, D.W. & Lindenbaum, M. (2005). On the equivalence of common approaches to lighting insensitive recognition, *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1721-1726.

Pantic, M. (2006). Face for Ambient Interface. *Lecture Notes in Artificial Intelligence*, vol. 3864, pp. 35-66.

Pantic, M. & Patras, I. (2005). Detecting facial actions and their temporal segments in nearly frontal-view face image sequences, *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, pp. 3358-3363.

Pantic, M. & Patras, I. (2006). Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. on Systems, Man and Cybernetics - Part B*, Vol. 36, No. 2, pp. 433-449.

Pantic, M., Pentland, A., Nijholt, A. & Huang, T. (2006). Human Computing and machine understanding of human behaviour: A Survey, *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 239-248.

Pantic, M. & Rothkrantz, L.J.M. (2000). Automatic Analysis of Facial Expressions – The State of the Art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, pp. 1424-1445.

Pantic, M. & Rothkrantz, L.J.M. (2003). Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE*, Spec. Issue on Human-Computer Multimodal Interface, Vol. 91, No. 9, pp. 1370-1390.

Pantic, M. & Rothkrantz, L.J.M. (2004a). Facial action recognition for facial expression analysis from static face images. *IEEE Trans. on Systems, Man and Cybernetics - Part B*, Vol. 34, No. 3, pp. 1449-1461.

Pantic, M. & Rothkrantz, L.J.M. (2004b). Case-based reasoning for user-profiled recognition of emotions from face images, *Proc. IEEE Int'l Conf. Multimedia & Expo*, pp. 391-394.

Pantic, M., Rothkrantz, L.J.M. & Koppelaar, H. (1998). Automation of non-verbal communication of facial expressions, *Proc. Conf. Euromedia*, pp. 86-93.

Pantic, M., Sebe, N., Cohn, J.F. & Huang, T. (2005a). Affective Multimodal Human-Computer Interaction, *Proc. ACM Int'l Conf. on Multimedia*, pp. 669-676.

Pantic, M., Valstar, M.F., Rademaker, R. & Maat, L. (2005b). Web-based database for facial expression analysis, *Proc. IEEE Int'l Conf. Multimedia and Expo*, pp. 317-321. (www.mmifacedb.com)

Patras, I. & Pantic, M. (2004). Particle filtering with factorized likelihoods for tracking facial features, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 97-102.

Patras, I. & Pantic, M. (2005). Tracking deformable motion, *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, pp. 1066-1071.

Pentland, A., Moghaddam, B. & Starner, T. (1994). View-Based and Modular Eigenspaces for Face Recognition, *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 84-91.

Picard, R.W. (1997). *Affective Computing*, MIT Press, Cambridge, USA.

Pitt, M.K. & Shephard, N. (1999). Filtering via simulation: auxiliary particle filtering. *J. Amer. Stat. Assoc.*, Vol. 94, pp. 590-599.

Rinn, W. E. (1984). The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, Vol. 95, No. 1, pp. 52–77.

Roark, D.A., Barret, S.E., Spence, M.J., Abdi, H. & O'Toole, A.J. (2003). Psychological and neural perspectives on the role of motion in face recognition. *Behavioral and cognitive neuroscience reviews*, Vol. 2, No. 1, pp. 15-46.

Rowley, H., Baluja, S. & Kanade, T. (1998). Neural Network-Based Face Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, pp. 23-38.

Russell, J.A. & Fernandez-Dols, J.M., (Eds.), (1997). *The Pshychology of Facial Expression*, Cambridge University Press, New York, USA.

Samal, A. & Iyengar, P.A. (1992). Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, Vol. 25, No. 1, pp. 65-77.

Sayette, M.A., Smith, D.W., Breiner, M.J. & Wilson, G.T. (1992). The effect of alcohol on emotional response to a social stressor. *Journal of Studies on Alcohol*, Vol. 53, pp. 541–545.

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

Schneiderman, H. & Kanade, T. (2000). A statistical model for 3D object detection applied to faces and cars, *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 746-751.

Shi, J. & Tomasi, C. (1994). Good features to track, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 593-600.

Su, C., Zhuang, Y., Huang, L. & Wu, F. (2004). A Two-Step Approach to Multiple Facial Feature Tracking: Temporal Particle Filter and Spatial Belief Propagation, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 433-438.

Sung, K.K. & Poggio, T. (1998). Example-Based Learning for View-Based Human Face Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, pp. 39-51.

Tao, H. & Huang, T.S. (1998). Connected vibrations – a model analysis approach to non-rigid motion tracking, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 735-740.

Tian, Y.L., Kanade, T. & Cohn, J.F. (2001). Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 97-115.

Tian, Y.L., Kanade, T. & Cohn, J.F. (2005). Facial Expression Analysis, In: *Handbook of Face Recognition*, Li, S.Z. & Jain, A.K., (Eds.), pp. 247-276, Springer, New York, USA.

Tong, Y., Liao, W. & Ji, Q. (2006). Inferring facial action units with causal relations, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 1623-1630.

Torres, E. & Andersen, R. (2006). Space-time separation during obstacle-avoidance learning in monkeys. *J. Neurophysiology*, Vol. 96, pp. 2613-2632.

Valstar, M.F. & Pantic, M. (2006a). Fully automatic facial action unit detection and temporal analysis, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, Vol. 3, p. 149.

Valstar, M.F. & Pantic, M. (2006b). Biologically vs. logic inspired encoding of facial actions and emotions in video, *Proc. IEEE Int'l Conf. Multimedia and Expo*.

Valstar, M.F., Pantic, M., Ambadar, Z. & Cohn, J.F. (2006). Spontaneous vs. posed facial behavior: Automatic analysis of brow actions, *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 162-170.

Valstar, M., Pantic, M. & Patras, I. (2004). Motion History for Facial Action Detection from Face Video, *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, Vol. 1, pp. 635-640.

Viola, P. & Jones, M. (2004). Robust real-time face detection. *J. Computer Vision*, Vol. 57, No. 2, pp. 137-154.

Vukadinovic, D. & Pantic, M. (2005). Fully automatic facial feature point detection using Gabor feature based boosted classifiers, *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, pp. 1692-1698.

Wang, P. & Ji, Q. (2004). Multi-View Face Detection under Complex Scene based on Combined SVMs, *Proc. IEEE Int'l Conf. Pattern Recognition*, Vol. 4, pp. 179-182.

Williams, A.C. (2002). Facial expression of pain: An evolutionary account. *Behavioral & Brain Sciences,* Vol. 25, No. 4, pp. 439-488.

Whitehill, J. & Omlin, C. 2006). Haar Features for FACS AU Recognition, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, 5 pp.

Xiao, J., Baker, S., Matthews, I. & Kanade, T. (2004). Real-time Combined 2D+3D Active Appearance Models, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 535-542.

Xiao, J., Moriyama, T., Kanade, T. & Cohn, J.F. (2003). Robust full-motion recovery of head by dynamic templates and re-registration techniques. *Int'l J. Imaging Systems and Technology,* Vol. 13, No. 1, pp. 85-94.

Yacoob, Y., Davis, L., Black, M., Gavrila, D., Horprasert, T. & Morimoto, C. (1998). Looking at People in Action, In: *Computer Vision for Human-Machine Interaction*, Cipolla, R. & Pentland, A., (Eds.), pp. 171-187, Cambridge University Press, Cambridge, UK.

Yang, M.H., Kriegman, D.J. & Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, pp. 34-58.

Yin, L., Wei, X., Sun, Y., Wang, J. & Rosato, M. (2006). A 3d facial expression database for facial behavior research, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 211-216.

Young, A.W., (Ed.), (1998). *Face and Mind*, Oxford University Press, Oxford, UK.

Zhang, Y. & Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequence. *IEEE Trans. Pattern Analysis & Machine Intelligence*, Vol. 27, No. 5, pp. 699-714.

Zhang, Z., Lyons, M., Schuster, M. & Akamatsu, S. (1998). Comparison Between Geometry-based and Garbor-Wavelet- based Facial Expression Recognition Using Multi-layer Perceptron, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 454-459.

Zhao, W., Chellappa, R., Rosenfeld, A. & Phillips, P.J. (2003). Face Recognition – A literature survey. *ACM Computing Surveys*, Vol. 35, No. 4, pp. 399-458.