PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Machine analysis of facial behaviour: naturalistic and dynamic behaviour

Maja Pantic

Phil. Trans. R. Soc. B 2009 **364**, 3505-3513
doi: 10.1098/rstb.2009.0135

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2009) **364**, 3505–3513
doi:10.1098/rstb.2009.0135

Machine analysis of facial behaviour: naturalistic and dynamic behaviour

Maja Pantic[1,2,*]

[1]*Department of Computing, Imperial College London, London SW7 2AZ, UK*
[2]*EEMCS, University of Twente, 7500 AE Enschede, The Netherlands*

This article introduces recent advances in the machine analysis of facial expressions. It describes the problem space, surveys the problem domain and examines the state of the art. Two recent research topics are discussed with particular attention: analysis of facial dynamics and analysis of naturalistic (spontaneously displayed) facial behaviour. Scientific and engineering challenges in the field in general, and in these specific subproblem areas in particular, are discussed and recommendations for accomplishing a better facial expression measurement technology are outlined.

Keywords: automatic facial expression analysis; dynamics of facial behaviour; naturalistic behaviour

1. INTRODUCTION

A widely accepted prediction is that computing will move to the background, weaving itself into the fabric of our everyday living and projecting the human user into the foreground. To realize this goal, next-generation computing (a.k.a. pervasive computing, ambient intelligence and human computing) will need to develop human-centred user interfaces that respond readily to naturally occurring, multimodal, human communication (Pantic *et al.* 2008). These interfaces will need the capacity to perceive and understand intentions and emotions as communicated by social and affective signals. Motivated by this vision of the future, automated analysis of non-verbal behaviour, and especially of facial behaviour, has attracted increasing attention in computer vision, pattern recognition and human–computer interaction (Pantic & Rothkrantz 2003; Pantic & Bartlett 2007; Vinciarelli *et al.* 2009; Zeng *et al.* 2009). To wit, facial expression is one of the most cogent, naturally preeminent means for human beings to communicate emotions, to clarify and stress what is said, and to signal comprehension, disagreement and intentions, in brief, to regulate interactions with the environment and other persons in the vicinity (Ambady & Rosenthal 1992; Ekman & Rosenberg 2005). Automatic analysis of facial expressions therefore forms the essence of numerous next-generation computing tools, including affective computing technologies (proactive and affective user interfaces), learner-adaptive tutoring systems, patient-profiled personal wellness technologies, etc.

This article introduces recent advances in the machine analysis of facial expressions. It describes the problem space, surveys the problem domain and examines the state of the art. Two recent research topics will receive particular attention: analysis of facial dynamics and analysis of naturalistic (spontaneously displayed) facial behaviour. Scientific and engineering challenges in the field in general, and in these specific subproblem areas in particular, will be discussed and recommendations for accomplishing a better facial expression measurement technology will be outlined.

*m.pantic@imperial.ac.uk

One contribution of 17 to a Discussion Meeting Issue 'Computation of emotions in man and machines'.

2. PROCESS OF AUTOMATIC FACIAL BEHAVIOUR ANALYSIS

Facial expression recognition is a process performed by humans or computers, which consists of three steps (figure 1):

(i) locating faces in the scene (e.g. in an image this step is also referred to as *face detection*),
(ii) extracting facial features from the detected face region (e.g. detecting the shape of facial components or describing the texture of the skin in a facial area; this step is referred to as *facial feature extraction*), and
(iii) analysing the motion of facial features and/or changes in the appearance of facial features and classifying this information into some facial-expression-interpretative categories such as facial muscle activations like smile or frown, emotion (affect) categories like happiness or anger, attitude categories like (dis)liking or ambivalence, etc. (this step is also referred to as *facial expression interpretation*).

The problem of *finding faces* can be viewed as a segmentation problem (in machine vision) or as a detection problem (in pattern recognition). It refers to identification of all regions in the scene that contain a human face. The problem of finding faces (*face localization*, *face detection*) should be solved regardless of clutter, occlusions and variations in head pose and lighting conditions. The presence of non-rigid movements owing to facial expression and a high degree of variability in facial size, colour and texture

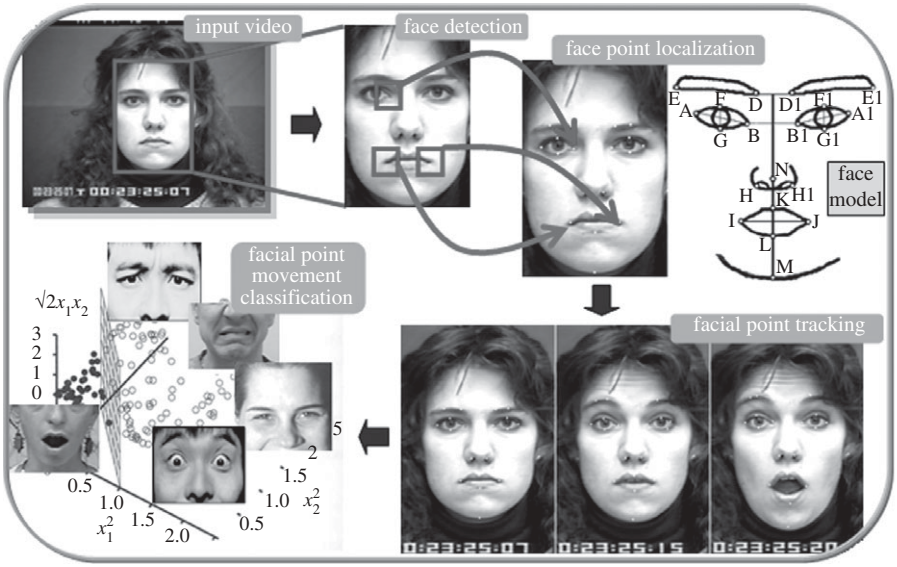

Figure 1. Outline of an automated, geometric-features-based system for facial expression recognition (for details of this system, see Valstar & Pantic 2007).

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. Arguably the most commonly employed face detector in automatic facial expression analysis is the real-time face detector proposed by Viola & Jones (2004).

The problem of feature extraction can be viewed as a dimensionality reduction problem (in machine vision and pattern recognition). It refers to transforming the input data into a reduced representation set of features, which encode the relevant information from the input data. The problem of *facial feature extraction* from input images may be divided into at least three dimensions (Pantic & Rothkrantz 2003; Pantic & Bartlett 2007): (i) Are the features holistic (spanning the whole face) or atomistic (spanning subparts of the face)? (ii) Is temporal information used? (iii) Are the features view based or volume based (two dimensional/three dimensional)? Given this glossary, most of the proposed approaches to facial expression recognition are directed towards static, analytic, two-dimensional facial feature extraction (Pantic & Bartlett 2007; Zeng *et al.* 2009). The usually extracted facial features are either *geometric features* such as the shapes of 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 in specific facial areas including wrinkles, bulges and furrows. Appearance-based features include learned image filters from independent component analysis (ICA), principal component analysis (PCA), local feature analysis (LFA), Gabor filters, integral image filters (also known as box filters and Haar-like filters), features based on edge-oriented histograms, etc. Several efforts have also been reported that use both geometric and appearance features (e.g. Zhang & Ji 2005). These approaches to automatic facial expression analysis are referred to as *hybrid* methods. Although 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, recent studies show that in some cases geometric features can outperform appearance-based ones (Pantic & Patras 2006; Pantic & Bartlett 2007). 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; Koelstra & Pantic 2008).

Contractions of facial muscles, which produce facial expressions, induce movements of the facial skin and changes in the location and/or appearance of facial features (e.g. contraction of the corrugator muscle induces a frown and causes the eyebrows to move towards each other, usually producing wrinkles between the eyebrows; figure 2). Such *changes can be detected* by analysing optical flow, facial-point- or facial-component-contour-tracking results, or by using an ensemble of classifiers trained to make decisions about the presence of certain changes (e.g. whether the nasolabial furrow is deepened or not) based on passed appearance features (see also §3). 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 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 facial expressions directly. Many researchers adopted this approach (e.g. Gokturk *et al.* 2002; Cohn *et al.* 2004; for comprehensive overviews, see Pantic & Rothkrantz 2000; Pantic & Rothkrantz 2003; Zeng *et al.* 2009). Until recently, standard optical flow techniques were arguably most commonly used for tracking facial characteristic points and contours as well. In order to address the limitations inherent in optical flow techniques such as the accumulation of error and the sensitivity to noise,

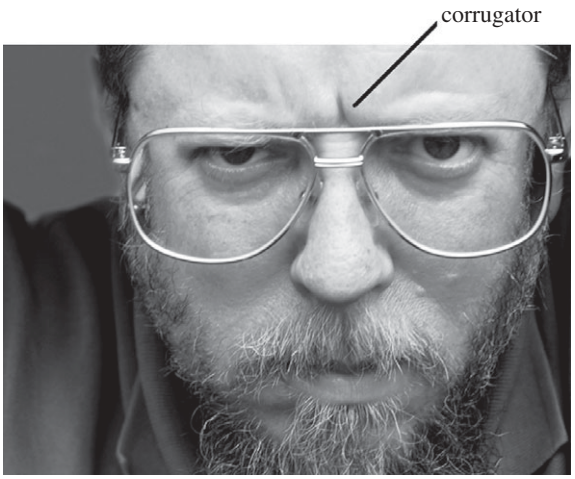

Figure 2. Facial appearance of the corrugator muscle contraction (coded as AU4 in the FACS system; Ekman *et al.* 2002).

occlusion, clutter and changes in illumination, recent efforts in automatic facial expression recognition use sequential state estimation techniques (such as Kalman filter and particle filter) to track facial feature points in image sequences (e.g. Zhang & Ji 2005; Pantic & Patras 2006; Pantic & Bartlett 2007).

Eventually, dense flow information, tracked movements of facial characteristic points, tracked changes in contours of facial components and/or extracted appearance features are translated into a description of the displayed facial expression. This description (*facial expression interpretation*) is usually given either in terms of shown affective states (emotions) or in terms of activated facial muscles underlying the displayed facial expression (see §3 for a detailed discussion and an overview of the state of the art). Most facial expression analysers 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 & Rothkrantz 2003; Vinciarelli *et al.* 2009). However, several promising prototype systems have been reported that can recognize deliberately produced action units (AUs) in face images, and even a few attempts towards recognition of spontaneously displayed AUs have been recently reported as well (Zeng *et al.* 2009; see also §4). While the older methods employ simple approaches, including expert rules and machine learning methods such as neural networks, to classify relevant information from the input data into some facial-expression-interpretative categories (Pantic & Rothkrantz 2000, 2003), the more recent (and often more advanced) methods employ probabilistic, statistical and ensemble learning techniques, which seem to be particularly suitable for automatic facial expression recognition from face image sequences (Pantic & Bartlett 2007; Zeng *et al.* 2009).

3. FACIAL BEHAVIOUR INTERPRETATION: EMOTIONS, SOCIAL SIGNALS AND ACTION UNITS

Two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) detection and facial muscle action (AU) detection (Pantic & Rothkrantz 2000, 2003; Zeng *et al.* 2009). These streams stem directly from two major approaches to facial expression measurement in psychological research (Cohn & Ekman 2005): message and sign judgement. The aim of message judgement is to infer what underlies a displayed facial expression, such as affect or personality, while the aim of sign judgement is to describe the 'surface' of the shown behaviour, such as facial movement or facial component shape. Thus, a brow frown (figure 2) can be judged as 'anger' in a message judgement and as a facial movement that lowers and pulls the eyebrows closer together in a sign-judgement approach. While message judgement is all about interpretation, sign judgement attempts to be objective, leaving inference about the conveyed message to higher order decision making.

The most commonly used facial expression descriptors in message-judgement approaches are the six basic emotions (fear, sadness, happiness, anger, disgust and surprise; figure 3), proposed by Ekman and discrete emotion theorists Keltner & Ekman (2000), who suggest that these emotions are universally displayed and recognized from facial expressions. This trend can also be found in the field of automatic facial expression analysis. Most facial expression analysers developed so far target human facial affect analysis and attempt to recognize a small set of prototypic emotional facial expressions like happiness and anger (Zeng *et al.* 2009). Automatic detection of the six basic emotions in posed, controlled displays can be done with reasonably high accuracy. More specifically, recent studies report a recognition accuracy of above 90 per cent for prototypic facial expressions of basic emotions displayed on command (e.g. Littlewort *et al.* 2006). Even though automatic recognition of acted expressions of the six basic emotions from face images and image sequences is considered largely solved, reports on novel approaches are published even now (e.g. Kotsia & Pitas 2007). Exceptions from this overall state of the art in the machine analysis of human facial affect include a few tentative efforts to detect acted expressions of cognitive and psychological states like interest (El Kaliouby & Robinson 2004), fatigue (Ji *et al.* 2006) and pain (Littlewort *et al.* 2007). 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 (see §4).

The most commonly used facial action descriptors in sign judgement-approaches are the AUs defined in the Facial Action Coding System (FACS; Ekman *et al.* 2002). FACS associates facial expression changes with actions of the muscles that produce them. It defines nine different AUs in the upper face, 18 in the lower face, five miscellaneous ones, 11 action descriptors (ADs) for head position, nine ADs for eye position and 14 additional descriptors for miscellaneous actions (figure 4). AUs are considered to be the smallest visually discernable facial movements. 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

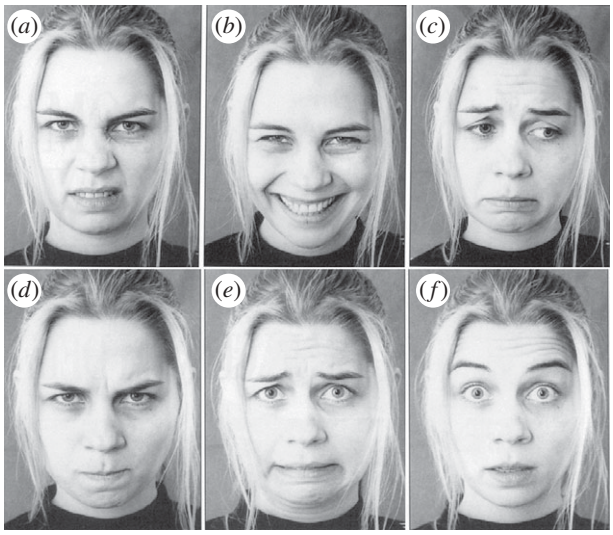

Figure 3. Prototypic facial expressions of six basic emotions (*a–f*): disgust, happiness, sadness, anger, fear and surprise.

anatomically possible facial expression, decomposing it into the specific AUs and their temporal segments that produced the expression. As AUs are independent of interpretation, they can be used for any higher order decision-making process, including recognition of basic emotions (based on EMFACS rules; Ekman *et al.* 2002), cognitive states like puzzlement (Cunningham *et al.* 2004), psychological states like suicidal depression (Ekman & Rosenberg 2005) or pain (Williams 2002), social behaviours like accord and rapport (Ambady & Rosenthal 1992; Cunningham *et al.* 2004), personality traits like extraversion and temperament (Ekman & Rosenberg 2005), 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; Ambady & Rosenthal 1992). 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 behaviours associated with cognitive states like comprehension, or social behaviours like empathy or politeness? How do we build systems to detect comprehension, for example, when we do not know for certain what faces display when students are comprehending? Having subjects pose mental states such as comprehension and puzzlement 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 (see also §4). Likewise, subjective labelling of expressions has also been shown to be less reliable than objective coding for finding relationships between facial expression and other state variables. An example where subjective judgements 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; Frank & Ekman 2004). Hence, AUs are very suitable for use as mid-level parameters in automatic facial behaviour analysis, as the thousands of anatomically possible expressions (Cohn & Ekman 2005) can be described as

Figure 4. Examples of AUs and their combinations defined in FACS.

combinations of 32 AUs and can be mapped to any higher order facial display interpretation, including basic emotions, cognitive states, social signals and behaviours, and complex mental states like depression.

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 (Pantic & Rothkrantz 2000). Several promising prototype systems that can recognize deliberately produced AUs in either (near-) frontal view face images (e.g. Bartlett *et al.* 1999; Tian *et al.* 2001; Pantic & Rothkrantz 2004) or profile view face images (Pantic & Rothkrantz 2004; Pantic & Patras 2006) were reported. These systems employ ranges of approaches, including expert rules, machine learning methods such as neural networks, feature-based image representations (i.e. using geometric features like facial points or shapes of facial components; see also §2) or appearance-based image representations (i.e. using texture of the facial skin including wrinkles and furrows; see also §2).

One of the main criticisms that this work has 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 behaviour rather than the exaggerated movements 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 pieces of work have recently emerged on machine analysis of AUs in spontaneous facial expression data. Section 4 provides a detailed discussion of these techniques and the challenges that face researchers of vision-based analysis of human naturalistic facial behaviour.

4. AUTOMATIC ANALYSIS OF FACIAL BEHAVIOUR: ACTED VERSUS NATURALISTIC BEHAVIOUR

The importance of making a clear distinction between spontaneous and deliberately displayed facial behaviour

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 facial actions originate in the subcortical areas of the brain. The facial expressions mediated by these two pathways have differences with respect to the way in which facial muscles are moved and in their dynamics (Ekman 2003; 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. For instance, it has been shown that spontaneous smiles, in contrast to posed smiles (e.g. a polite smile), have smoother transitions between onset, apex and offset of movement (Frank *et al.* 1993), can have multiple AU12 apexes (multiple rises of the mouth corners) and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1 s (Cohn & Schmidt 2004). 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 AUs, their temporal segments and intensity frame-by-frame (the manual coding of 1 min of video tape takes approx. 1 h). Hence the importance of video-based automatic coding systems.

Nonetheless, as already mentioned above, most of the existing work on automatic facial expression recognition is based on deliberate and often exaggerated facial expressions (for survey papers on past work in the field, see Pantic & Rothkrantz 2000, 2003; Zeng *et al.* 2009). Little work has been reported on the machine analysis of spontaneous facial expression data. When it comes to automatic recognition of affective and mental states from naturalistic facial behaviour data, a few tentative efforts have been reported to detect naturalistic expressions of cognitive and psychological states like frustration (Kapoor *et al.* 2007), fatigue (Fan *et al.* 2008) and pain (Ashraf *et al.* 2007; Littlewort *et al.* 2007). Also, a few studies investigating the dimensional approach to automatic affect recognition have been reported. For example, the study by Zeng *et al.* (2006) investigated automatic discrimination between positive and negative affect, while the study by Ioannou *et al.* (2005) investigated classification of input facial expression data into quadrants in evaluation–activation space. A small number of studies has also been reported on the automatic recognition of AUs in naturalistic facial behaviour data. These include studies on upper-face AUs only (Kapoor *et al.* 2003; Cohn *et al.* 2004; Valstar *et al.* 2006) as well as on all AUs (Bartlett *et al.* 2005). Finally, several recent studies explicitly investigated the difference between spontaneous and deliberate facial behaviour. The work by Valstar *et al.* (2006) 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. The work by Littlewort *et al.* from 2007 (Bartlett *et al.* 2005) reports on an automated system for discriminating genuine from faked pain facial expressions. The system was built around morphological (rather than temporal) characteristics of genuine pain expression (i.e. presence of certain AUs and their intensity). The work by Valstar *et al.* from 2007 (Valstar *et al.* 2007) concerns an automated system for distinguishing acted from spontaneous smiles. The study shows that combining information from multiple visual cues (in this case, facial expressions, head movements and shoulder movements) outperforms single-cue approaches to the target problem. It also clearly shows that the differences between spontaneous and deliberately displayed smiles are in the dynamics of shown behaviour (e.g. amount of head and shoulder movement, the speed of onset and offset of the actions, and order and timing of the actions' occurrences) rather than in the configuration of the displayed expression. These findings are in accordance with the research findings in psychology (e.g. Cohn & Schmidt 2004; Krumhuber *et al.* 2007). Most of the existing systems for facial expression analysis in naturalistic data are based on two-dimensional spatial or spatio-temporal facial features and employ advanced probabilistic (e.g. coupled and triple hidden Markov models (HMM)), statistical (e.g. support vector machines (SVM) and relevance vector machines (RVM)) and ensemble learning techniques (e.g. Adaboost and Gentleboost).

Although it is obvious that methods of automated facial behaviour analysis that have been trained on deliberate and often exaggerated behaviours may fail to generalize to the complexity of expressive behaviour found in real-world settings (and most probably will fail given the fact that deliberate behaviour differs in visual appearance and timing from spontaneously occurring behaviour), relatively few efforts have been reported so far towards the development of systems trained and tested on naturalistic behaviour. There are at least two reasons for this. Firstly, automatic analysis of spontaneously occurring behaviour can hardly be done without analysing the dynamics of the displayed behaviour, which, in turn, is a barely investigated research topic as explained in §5. Secondly, to develop and evaluate facial behaviour analysers capable of dealing with spontaneously occurring behaviour, large collections of suitable, annotated, publicly available training and test data are needed, which, currently, is not the case (see §6 for a further discussion on this topic).

5. AUTOMATIC ANALYSIS OF FACIAL BEHAVIOUR: DYNAMICS OF FACIAL BEHAVIOUR

Automatic recognition of facial expression configuration (in terms of AUs constituting the observed

expression) has been the main focus of research efforts in the field. However, both the configuration and the dynamics of facial expressions (i.e. timing, duration, speed of activation and deactivation of various AUs, etc.) are important for the interpretation of human facial behaviour. The body of research in cognitive sciences, which argues that the dynamics of facial expressions is crucial for the interpretation of observed behaviour, is ever growing (Russell & Fernandez-Dols 1997; Ambadar *et al.* 2005; Ekman & Rosenberg 2005). Facial expression temporal dynamics is essential for the categorization of complex psychological states like various types of pain and mood (Williams 2002). They improve the judgement of observed facial behaviour (e.g. affect) by enhancing the perception of change and by facilitating the processing of facial configuration (Ambadar *et al.* 2005). They represent a critical factor for interpretation of social behaviours like social inhibition, embarrassment, amusement and shame (Costa *et al.* 2001; Ekman & Rosenberg 2005). They are also a key parameter in differentiating between posed and spontaneous facial displays (Frank *et al.* 1993; Ekman 2003; Cohn & Schmidt 2004; Frank & Ekman 2004), as explained in §4.

In spite of these findings, the vast majority of past work on machine analysis of human facial behaviour does not take the dynamics of facial expressions into account when analysing shown facial behaviour. Some of the past work in the field has used aspects of the temporal dynamics of facial expression such as the speed of a facial point displacement or the persistence of facial parameters over time. However, this was mainly done in order to increase the performance of facial expression analysers (e.g. Zhang & Ji 2005; Gralewski *et al.* 2006; Tong *et al.* 2007) or to report on the intensity of (a component of) the shown facial expression (e.g. Zhang & Ji 2005; Littlewort *et al.* 2006) rather than to explicitly encode the temporal dynamics of shown facial behaviour. Only a few recent studies analyse explicitly the temporal dynamics of facial expressions. These studies explore feature-based approaches to automatic segmentation of AU activation into temporal segments (neutral, onset, apex, offset) in frontal-view (Pantic & Patras 2005; Valstar & Pantic 2007) and profile-view (Pantic & Patras 2006) face videos, appearance-based approaches to automatic coding of temporal segments of AUs (Koelstra & Pantic 2008) and approaches to modelling temporal relationships between AUs as present in expressions of basic emotions (Tong *et al.* 2007).

The work by Pantic & Patras (2005, 2006) employs rule-based reasoning to encode AUs and their temporal segments based on a set of spatiotemporal features extracted from the trajectories of tracked facial characteristic points. 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 behaviour' (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.* 2005*a*). Experimental evidence supporting this assumption for the case of prototypic emotional facial expressions was reported in Valstar & Pantic (2006). Experimental evidence supporting this assumption for the case of expression configuration detection and its temporal activation model (neutral → onset → apex → offset) recognition has been recently reported as well (Valstar & Pantic 2007). In this latter work, a number of facial characteristic points are detected and tracked in an input face video (the particle filtering framework has been used for tracking purposes), a set of spatiotemporal features is extracted from the trajectories of the tracked points, and a combination of statistical and probabilistic machine learning techniques (namely a combination of SVM and HMM) is used to detect AUs and their temporal segments (see figure 1 for the outline of the method). The reported experimental results clearly show that modelling facial expression temporal dynamics and analysing displayed facial expressions based on such models significantly improve the performance of the automated system (an increase of 6% in terms of the F_1 measure was reported).

The work by Koelstra & Pantic (2008) proposes an appearance-based approach to automatic coding of AUs and their temporal segments. It presents a dynamic-texture-based approach based on non-rigid registration using free-form deformations, in which the extracted facial motion representation is used to derive motion orientation histogram descriptors in both the spatial and temporal domain, which, in turn, form further input to a set of AU classifiers based on ensemble and probabilistic machine learning techniques (more specifically, a combination of Gentleboost and HMM was used). This work represents the first appearance-based approach to explicit segmentation of AU activation into temporal segments, reconfirming the results reported in Valstar & Pantic (2007)—modelling facial expression temporal dynamics and analysing displayed facial expressions based on such models significantly improve the performance of the automated system.

The only work reported so far on modelling the temporal correlation of different AUs is that by Tong *et al.* (2007). It applies the appearance-based approach to AU recognition, similar to that by Littlewort *et al.* (2006), using Gabor features and a set of Gentleboost classifiers, one for each target AU. Furthermore it uses a hierarchical probabilistic framework (more specifically, dynamic Bayesian networks) to model the relationships among different AUs as found in facial expressions of the six basic emotions. The work reconfirms the results reported in Valstar & Pantic (2007)—the integration of AU relationships and AU dynamics with AU measurements yields a significant improvement of AU recognition (an increase of 5% in the correct recognition rate was reported).

Although these pioneering efforts towards automatic analysis of the temporal structure of facial expressions are truly promising, many research issues are open and yet to be investigated. A crucial issue that remains unresolved is how the grammar of naturalistic facial behaviour can be learned and how this information can be properly represented and used to handle ambiguities in the input data. Another important issue relates to multi-cue visual analysis. Except for a few studies (e.g. Cohn *et al.* 2004; Valstar *et al.* 2007), existing efforts towards the machine analysis of facial behaviour focus only on the analysis of facial gestures without taking into consideration other visual cues such as head movements, gaze patterns and body gestures like shoulder movements. However, research in cognitive science reports that human judgements of behavioural cues are the most accurate when both the face and the body are taken into account (Ambady & Rosenthal 1992). Experimental evidence supporting this finding for the case of automatic laughter analysis was reported in Valstar *et al.* (2007). Taking into account both face and body movements seems to be of particular importance when judging certain complex mental states such as embarrassment (Costa *et al.* 2001). However, integration, temporal structures and temporal correlations between different visual cues are virtually unexplored areas of research.

6. EVALUATING THE PERFORMANCE OF AN AUTOMATED SYSTEM FOR FACIAL BEHAVIOUR ANALYSIS

The final step in the development of automated systems for facial behaviour analysis is the performance analysis of a developed system. The two crucial aspects of evaluating the performance of a designed system are the used training/test dataset and the adopted evaluation strategy.

Having enough labelled data of the target human facial behaviour is a prerequisite in designing robust automatic facial expression recognizers. Explorations of this issue showed that, given an accurate three-dimensional alignment of the face, at least 50 training examples are needed for moderate performance (in the 80% accuracy range) of a machine learning approach to recognition of a specific facial expression (Pantic & Bartlett 2007). Recordings of spontaneous facial behaviour are difficult to collect because they are difficult to elicit, short lived and filled with subtle context-based changes. In addition, manual labelling of spontaneous facial behaviour for ground truth is very time consuming, error prone and expensive. Owing to these difficulties, most of the existing studies on automatic facial expression recognition are based on the 'artificial' material of deliberately displayed facial behaviour (see also §4), elicited by asking the subjects to perform a series of facial expressions in front of a camera. The most commonly used, publicly available, annotated datasets of posed facial expressions include the Cohn–Kanade facial expression database (Kanade *et al.* 2000) and the MMI facial expression database (Pantic *et al.* 2005*b*). Yet, as increasing evidence suggests that deliberate (posed) behaviour differs in appearance and timing from that which occurs in daily life (see §4), it is not surprising that approaches that have been trained on deliberate and often exaggerated behaviours usually fail to generalize to the complexity of expressive behaviour found in real-world settings. To address the general lack of a reference set of (audio and/or) visual recordings of human spontaneous behaviour, several efforts aimed at the development of such datasets have been recently reported. The most commonly used, publicly available, annotated datasets of spontaneous human behaviour recordings include the SAL dataset, the UT Dallas dataset and the MMI-part2 database (Pantic & Bartlett 2007; Zeng *et al.* 2009).

In pattern recognition and machine learning, a common evaluation strategy is to consider the correct classification rate (*classification accuracy*) or its complement error rate. However, this assumes that the natural distribution (prior probabilities) of each class is known and balanced. In an imbalanced setting, where the prior probability of the positive class is significantly less than the negative class (the ratio of these being defined as the *skew*), accuracy is inadequate as a performance measure since it becomes biased towards the majority class. That is, as the skew increases, accuracy tends towards majority class performance, effectively ignoring the recognition capability with respect to the minority class. This is a very common (if not the default) situation in a facial expression recognition setting, where the prior probability of each target class (a certain facial expression) is significantly less than the negative class (all other facial expressions). Thus, when evaluating the performance of an automatic facial expression recognizer, other performance measures such as *precision* (this indicates the probability of correctly detecting a positive test sample and is independent of class priors), *recall* (this indicates the fraction of positives detected that are actually correct and, as combines results from both positive and negative samples, it is class prior dependent), F_1*-measure* (this is calculated as $2 \times recall \times precision/(recall + precision)$) and ROC (this is calculated as $P(x|positive)/P(x|negative)$, where $P(x|C)$ denotes the conditional probability that a data entry has the class label C, and where a ROC curve plots the classification results from the most positive to the most negative classification) are more appropriate. However, because a confusion matrix shows all of the information about a classifier's performance, it should be used whenever possible for presenting the performance of the evaluated facial expression recognizer.

7. CONCLUDING REMARK

Faces are tangible projector panels of the mechanisms that govern our emotional and social behaviours. The automation of the entire process of facial behaviour analysis is, therefore, a highly intriguing problem, the solution to which would be enormously beneficial for fields as diverse as medicine, law, communication, education and computing. Although research in the field has seen a lot of progress in the past few years, several issues remain unresolved. Arguably the most

important unattended aspect of the problem is how the grammar of facial behaviour (i.e. temporal evolution of occurrences of visual cues including facial gestures, gaze patterns and body gestures like head and shoulder movements) can be learned and how this information can be properly represented and used to handle ambiguities in the observation data. This aspect of machine analysis of facial behaviour forms the main focus of the current and future research in the field.

The work of Maja Pantic is funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB) and in part by the European Community's 7th Framework Programme (FP7/2007–2013) under the grant agreement no. 231287 (SSPNet).

REFERENCES

Ambadar, Z., Schooler, J. & Cohn, J. F. 2005 Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions. *Psychol. Sci.* **16**, 403–410. (doi:10.1111/j.0956-7976.2005.01548.x)

Ambady, N. & Rosenthal, R. 1992 Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychol. Bull.* **111**, 256–274. (doi:10.1037/0033-2909.111.2.256)

Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K., Solomon, P. & Theobald, B. J. 2007 The painful face: pain expression recognition using active appearance models. *Proc. ACM Int. Conf. on Multimodal Interfaces*, 9–14.

Bartlett, M. S., Viola, P. A., Sejnowski, T. J., Golomb, B. A., Larsen, J., Hager, J. C. & Ekman, P. 1996 Classifying facial actions. *Adv. Neural Inf. Process. Syst.* **8**, 823–829.

Bartlett, M. S., Hager, J. C., Ekman, P. & Sejnowski, T. J. 1999 Measuring facial expressions by computer image analysis. *Psychophysiology* **36**, 253–263. (doi:10.1017/S0048577299971664)

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

Cohn, J. F. & Ekman, P. 2005 Measuring facial actions. In *The new handbook of methods in nonverbal behavior research* (eds J. A. Harrigan, R. Rosenthal & K. Scherer), pp. 9–64. New York, NY: Oxford University Press.

Cohn, J. F. & Schmidt, K. L. 2004 The timing of facial motion in posed and spontaneous smiles. *J. Wavelets Multi-Resolut. Inf. Process.* **2**, 121–132.

Cohn, J. F., Reed, L. I., Ambadar, Z., Xiao, J. & Moriyama, T. 2004 Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. *Proc. IEEE Int. Conf. Syst. Man Cybern.* **1**, 610–616.

Costa, M., Dinsbach, W., Manstead, A. S. R. & Bitti, P. E. R. 2001 Social presence, embarrassment, and nonverbal behaviour. *J. Nonverbal Behav.* **25**, 225–240. (doi:10.1023/A:1012544204986)

Cunningham, D. W., Kleiner, M., Wallraven, C. & Bülthoff, H. H. 2004 The components of conversational facial expressions. *Proc. ACM Int. Conf. on Applied Perception in Graphics and Visualization*, 143–149.

Ekman, P. 2003 Darwin, deception, and facial expression. *Ann. N. Y. Acad. Sci.* **1000**, 205–221. (doi:10.1196/annals.1280.010)

Ekman, P. & Friesen, W. V. 1969 The repertoire of nonverbal behavior. *Semiotica* **1**, 49–98.

Ekman, P. & Rosenberg, E. L. (eds) 2005 *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system*. Oxford, UK: Oxford University Press.

Ekman, P., Friesen, W. V. & Hager, J. C. 2002, *Facial action coding system*. Salt Lake City, UT: A Human Face.

El Kaliouby, R. & Robinson, P. 2004 Real-time inference of complex mental states from facial expressions and head gestures. *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.* **3**, 154.

Fan, X., Sun, Y. & Yin, B. 2008 Multi-scale dynamic human fatigue detection with feature level fusion. *Proc. IEEE Int. Conf. on Automatic Face & Gesture Recognition*.

Frank, M. G. & Ekman, P. 2004 Appearing truthful generalizes across different deception situations. *J. Pers. Soc. Psychol.* **86**, 486–495. (doi:10.1037/0022-3514.86.3.486)

Frank, M. G., Ekman, P. & Friesen, W. V. 1993 Behavioural markers and recognizability of the smile of enjoyment. *J. Pers. Soc. Psychol.* **64**, 83–93. (doi:10.1037/0022-3514.64.1.83)

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

Gralewski, L., Campbell, N. & Voak, I. P. 2006 Using a tensor framework for the analysis of facial dynamics. *Proc. IEEE Int. Conf. on Face & Gesture Recognition*, pp. 217–222.

Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T., Karpouzis, K. & Kollias, S. 2005 Emotion recognition through facial expression analysis based on a neurofuzzy method. *J. Neural Netw.* **18**, 423–435. (doi:10.1016/j.neunet.2005.03.004)

Ji, Q., Lan, P. & Looney, C. 2006 A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Trans. Syst. Man Cybern. A* **36**, 862–875. (doi:10.1109/TSMCA.2005.855922)

Kanade, T., Cohn, J. F. & Tian, Y. 2000 Comprehensive database for facial expression analysis. *Proc. IEEE Int. Conf. on Automatic Face & Gesture Recognition*, pp. 46–53.

Kapoor, A., Qi, Y. & Picard, R. W. 2003 Fully automatic upper facial action recognition. *Proc. IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures*.

Kapoor, A., Burleson, W. & Picard, R. W. 2007 Automatic prediction of frustration. *J. Hum. Comput. Stud.* **65**, 724–736.

Keltner, D. & Ekman, P. 2000 Facial expression of emotion. In *Handbook of emotions* (eds M. Lewis & J. M. Haviland-Jones), pp. 236–249. New York, NY: Guilford Press.

Koelstra, S. & Pantic, M. 2008 Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*.

Kotsia, I. & Pitas, I. 2007 Facial expression recognition in image sequences using geometric deformation features and SVM. *IEEE Trans. Image Process.* **16**, 172–187. (doi:10.1109/TIP.2006.884954)

Krumhuber, E., Manstead, A. S. R. & Kappas, A. 2007 Temporal aspects of facial displays in person and expression perception: the effects of smile dynamics, head-tilt, and gender. *J. Nonverbal Behav.* **31**, 39–56. (doi:10.1007/s10919-006-0019-x)

Li, S. Z. & Jain, A. K. (eds) 2005 *Handbook of face recognition*. New York, NY: Springer.

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. Conf. on Computer Vision & Pattern Recognition*, 853–859.

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J. & Movellan, J. 2006 Dynamics of facial expression extracted automatically from video. *J. Image Vis. Comput.* **24**, 615–625. (doi:10.1016/j.imavis.2005.09.011)

Littlewort, G. C., Bartlett, M. S. & Lee, K. 2007 Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. *Proc. ACM Int. Conf. on Multimodal Interfaces*, pp. 15–21.

Pantic, M. & Bartlett, M. S. 2007 Machine analysis of facial expressions. In *Face recognition* (eds K. Delac & M. Grgic), pp. 377–416. Vienna, Austria: I-Tech Education and Publishing.

Pantic, M. & Patras, I. 2005 Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. *Proc. IEEE Int. Conf. on Systems, Man & Cybernetics*, pp. 3358–3363. (doi:10.1109/TSMCB.2005.859075)

Pantic, M. & Patras, I. 2006 Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Syst. Man Cybern. B* **36**, 433–449.

Pantic, M. & Rothkrantz, L. J. M. 2000 Automatic analysis of facial expressions: the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1424–1445. (doi:10.1109/34.895976)

Pantic, M. & Rothkrantz, L. J. M. 2003 Toward an affect-sensitive multimodal HCI. *Proc. IEEE* **91**, 1370–1390. (doi:10.1109/JPROC.2003.817122)

Pantic, M. & Rothkrantz, L. J. M. 2004 Facial action recognition for facial expression analysis from static face images. *IEEE Trans. Syst. Man Cybern. B* **34**, 1449–1461. (doi:10.1109/TSMCB.2004.825931)

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

Pantic, M., Sebe, N., Cohn, J. F. & Huang, T. 2005*a* Affective multimodal human–computer interaction. *Proc. ACM Int. Conf. on Multimedia*, pp. 669–676

Pantic, M., Valstar, M. F., Rademaker, R. & Maat, L. 2005*b* Web-based database for facial expression analysis. *Proc. IEEE Int. Conf. on Multimedia & Expo*, pp. 317–321.

Pantic, M., Pentland, A., Nijholt, A. & Huang, T. S. 2008 Human-centred intelligent human–computer interaction (HCI^2): how far are we from attaining it? *J. Autonom. Adaptive Commun. Syst.* **1**, 168–187. (doi:10.1504/IJAACS.2008.019799)

Russell, J. A. & Fernandez-Dols, J. M. (eds) 1997 *The psychology of facial expression*. New York, NY: Cambridge University Press.

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

Tian, Y. L., Kanade, T. & Cohn, J. F. 2001 Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 97–115. (doi:10.1109/34.908962)

Tong, Y., Liao, W. & Ji, Q. 2007 Facial action unit recognition by exploiting their dynamics and semantic relationships. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 1683–1699. (doi:10.1109/TPAMI.2007.1094)

Valstar, M. F. & Pantic, M. 2006 Biologically vs. logic inspired encoding of facial actions and emotions in video. *Proc. IEEE Int. Conf. on Multimedia & Expo*, pp. 325–328.

Valstar, M. F. & Pantic, M. 2007 Support vector machines and hidden Markov models for modeling facial action temporal dynamics. *Lect. Notes Comput. Sci.* **4796**, 118–127. (doi:10.1007/978-3-540-75773-3_13)

Valstar, M. F., Pantic, M., Ambadar, Z. & Cohn, J. F. 2006 Spontaneous versus posed facial behavior: automatic analysis of brow actions. *Proc. ACM Int. Conf. on Multimodal Interfaces*, pp. 162–170.

Valstar, M. F., Gunes, H. & Pantic, M. 2007 How to distinguish posed from spontaneous smiles using geometric features. *Proc. ACM Int. Conf. on Multimodal Interfaces*, pp. 38–45.

Vinciarelli, A., Pantic, M. & Bourlard, H. 2009 Social signal processing: survey of an emerging domain. *J. Image Vis. Comput.* **27**, 1743–1760.

Viola, P. & Jones, M. 2004 Robust real-time face detection. *J. Comput. Vis.* **57**, 137–154. (doi:10.1023/B:VISI.0000013087.49260.fb)

Williams, A. C. 2002 Facial expression of pain: an evolutionary account. *Behav. Brain Sci.* **25**, 439–488.

Yang, M. H., Kriegman, D. J. & Ahuja, N. 2002 Detecting faces in images: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 34–58. (doi:10.1109/34.982883)

Zeng, Z., Fu, Y., Roisman, G. I., Wen, Z., Hu, Y. & Huang, T. S. 2006 Spontaneous emotional facial expression detection. *J. Multimedia* **1**, 1–8.

Zeng, Z., Pantic, M., Roisman, G. I. & Huang, T. S. 2009 A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 39–58. (doi:10.1109/TPAMI.2008.52)

Zhang, Y. & Ji, Q. 2005 Active and dynamic information fusion for facial expression understanding from image sequence. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 699–714. (doi:10.1109/TPAMI.2005.93)