Communication and Automatic Interpretation of Affect from Facial Expressions

Salah, A.A.; Sebe, N.; Gevers, T.

Publication date
2010
Document Version
Submitted manuscript
Published in
Affective computing and interaction: psychological, cognitive, and neuroscientific perspectives

Communication and automatic interpretation of affect from facial expressions[1]

Albert Ali Salah
University of Amsterdam, the Netherlands
Nicu Sebe
University of Trento, Italy
Theo Gevers
University of Amsterdam, the Netherlands

ABSTRACT

The objective of this chapter is to introduce the reader to the recent advances in computer processing of facial expressions and communicated affect. Human facial expressions have evolved in tandem with human face recognition abilities, and show remarkable consistency across cultures. Consequently, it is rewarding to review the main traits of face recognition in humans, as well as consolidated research on the categorization of facial expressions. The bulk of the chapter focuses on the main trends in computer analysis of facial expressions, sketching out the main algorithms and exposing computational considerations for different settings. We then look at some recent applications and promising new projects to give the reader a realistic view of what to expect from this technology now and in near future.

INTRODUCTION

In June 2009, Microsoft released a trailer of its latest project for Xbox 360 gaming console, called Project Natal. The video, an instant Facebook epidemic and a YouTube favourite, featured Peter Molyneux, the creative director of Microsoft Game Studios Europe, demonstrating a virtual agent called Milo. Using the sensing and processing capabilities of its hardware, the virtual agent communicated with the user as if the boundary of the screen is just a window, recognizing identity and speech, but also emotions, which enabled it to respond to the user with an impressive range of realistic behaviours. The innovation of the project was in its ambitious scope: creating a virtual agent that truly communicates with the user. The key to life–like communication was recognizing emotions of the user, and in return, generating states that carry affect information for the agent in human–readable form, i.e. in the body posture, vocal intonation, and most importantly, facial expression.

The recently flourishing field of social signal processing (Vinciarelli et al., 2009) targets a greater contextual awareness for computer systems and human–machine interaction, and drawing on cognitive psychology, places great emphasis on automatically understanding facial expressions. The human face is a window that allows peeking into diverse patterns of emotions that manifest themselves voluntarily and involuntarily, communicating affect or projected displays of personality. Even dissociated from gesture and voice (as in still face pictures), facial expressions convey complex, layered, and vital information. Consequently, it is a great challenge to create computer systems that can automatically analyse images to

[1] This is the uncorrected author proof. The copyright of this work rests with IGI Global. The full citation is Salah, A.A., N. Sebe, T. Gevers, "Communication and automatic interpretation of affect from facial expressions," in D. Gokcay & G. Yildirim (eds), Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives, IGI Global, (to appear, check http://staff.science.uva.nl/~asalah/ for updates on this edited book).

reveal the sometimes obvious and sometimes subtle messages engraved in faces. In this chapter we aim to provide the reader with a broad overview of how computer scientists have risen to this challenge, starting from relevant taxonomies and guidelines, briefly touching upon the cognitive aspects of affect and face recognition, summarizing recent advances in algorithmic aspects of the problem, giving pointers and tools for the initiate, and finally, discussing applications and the future of facial expression recognition.

CATEGORIZATION OF FACIAL EXPRESSIONS

The human face is a complicated visual object; it contains a lot of information with regards to identity, communicative intent and affect, and humans can "read" these cues, even under difficult visibility conditions. We can for instance understand the emotions of a person we see for the first time. In this section we look at taxonomies of facial expressions, and point out to several important factors that need to be taken into account in evaluating facial expressions.

A facial expression can be the result of an emotional response (spontaneous), or a construct with communicative intent (volitional) (Russell & Fernandez–Dols, 1997). It can occur naturally, or it can be posed. In both cases, it can have different intensities, and it can be a mixture of pure expressions. These factors make the task of sorting out a facial expression difficult. Additionally, the categorization of expressions can be achieved in ever-finer levels. It is one thing to label the category of an expression as "happy", quite another to distinguish between a real smile (also called a Duchenne smile), a miserable smile, an angry smile, an embarrassed smile, and a dimpler. Finally, cultural differences in facial expressions also need to be taken into account.

Categorization of emotions predate computers by hundreds of years, but the roles of particular emotions in society are different for each culture; in India, for instance, it was believed that the basic emotions are sexual passion, anger, disgust, perseverance, amusement, sorrow, wonder, fear, and serenity. Facial expressions of these emotions are culture-dependent, but also the semantic counterparts of these emotions do not completely overlap with the current understanding of these words, adding to the difficulty of systematically categorizing emotions. Furthermore, the experimental settings under which any study is conducted and the ensuing databases on which we measure the success of a given method are not independent of cultural influences. For instance it is known that in some cultures the expression of emotion is more restricted for social reasons. Finally, as facial morphology also changes according to the anthropological group of a subject, it is natural to expect some principled variation across races and gender.

Adolphs (2002) cautions that the emotion categories used in everyday life may not result in the most appropriate categorization of emotion for scientific purposes. Yet, as opposed to language, automatic perceptual grouping of emotional cues is apt to produce meaningful structural relationships, as the semantic proximity of emotions is reflected in the structural proximity of their expressions. Presently, most emotion researchers use discrete categories to indicate the presence of one emotion in any given instance. This method requires singular labelling of the ground truth, as well as mostly exaggerated expressions in the evaluation data. Paul Ekman argued that **six basic emotional facial expression categories** are persistent across cultures (Ekman, 1993): happiness, sadness, anger, fear, surprise and disgust. While other taxonomies extend these basic emotions with contempt, shame, distress, interest, guilt, and many others, this six–emotion classification is the most commonly used taxonomy in the computer science literature pertaining to facial expression analysis.

We note here that the presence of emotion, even when measured in continuous scales, is usually seen as a momentary evaluation of the percept, instead of a spatio–temporal event unfolding in time. A more granular analysis must be able to label subordinate components of a complex emotion, and the research is headed in this direction. One challenge is to create complex ground truth to measure methods that will attempt to gauge the accuracy of such an analysis.

The range of facial expressions is assumed to reflect the range of emotional displays in general, which is the primary reason we base facial expression categorization on taxonomies of emotional display. One important taxonomy is due to Russell (1980), and it dissects emotions that give rise to expressions along *arousal* and *valence* dimensions. Arousal controls the intensity of the emotion, whereas valence relates to its positive and negative connotations. This model enables a two-dimensional projection of emotional displays (See Figure 1). However, given a face image, it is difficult to precisely situate it in this projection. Note that of the six basic categories proposed by Ekman, only happiness has a positive valence. Another relevant approach is the OCC model (Ortony, Clore, & Collins, 1988), which proposes valenced reactions to consequences of events, actions of agents, and aspects of objects, based on relevant attributes thereof. Complex emotional theories usually take action semantics into account, which stresses the dynamic nature of emotions. For computer analysis, the incorporation of semantics is usually much more difficult due knowledge acquisition and representation issues. Consequently, computer scientists prefer to work with quantifiable schemes.

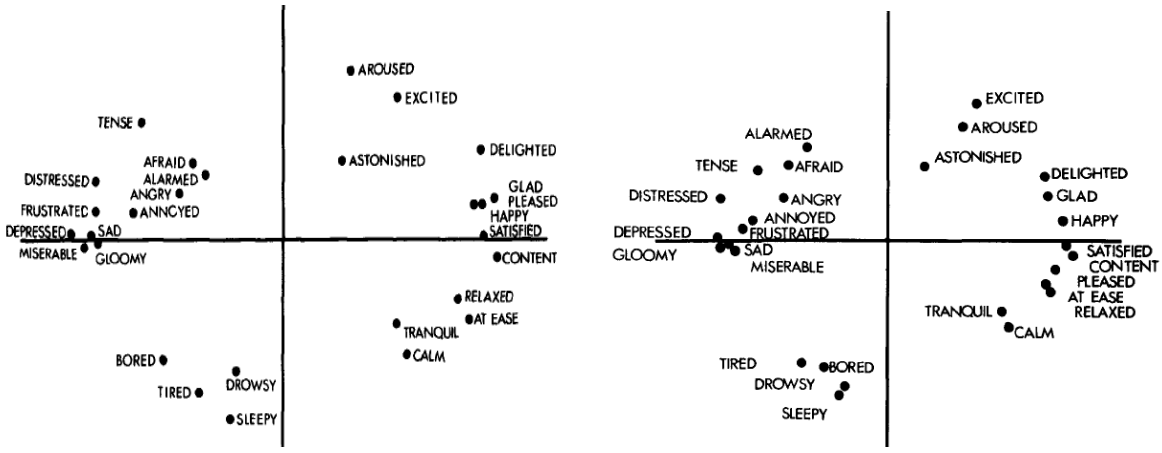


Figure 1. Arousal- and valence-based categorization of emotions agrees with self-report studies. Here are 28 affect words distributed according to valence/arousal (left) and the same words distributed by a principal components analysis of self-reports (right). © 1980, Russell. Used with permission.

The **facial action coding system** (FACS) developed by Ekman and Friesen (1978) is the basis of much work on facial expression recognition today. According to this system, facial muscle movements are grouped into different action units (AU), and expressions are described in terms of these action units. Each action is attributed with an intensity value (on a 5–point scale). There are procedures for describing the timing of facial actions, as well as for describing expressions as events. Table 1 gives a revised list of AUs. Historically, AU 40 is indicated as optional, and AUs 41 and 42 are merged into 43 ("Eyes Closure", previously "Eyes Closed") in later versions. References to FACS recognizing 44 action units refer to the old version, which omits AUs 3 and 8. AU 3 is the "Brow Knit", which is later dropped from the classification, and AU 8 was re-inserted. One of the reasons of the popularity of FACS is that facial muscle movements are completely objective physiological measures, and thus provide a solid basis for the formulation of emotion categories. Also, they are not restricted to displays of emotion, which increases their usability.

For other observational coding schemes and their comparison with FACS, the reader is referred to (Ekman & Rosenberg, 2005). We note here only that the FACS model only describes changes in facial configuration, and leaves temporal dynamics of these changes out of the picture. For relating facial expressions to emotions, Ekman and Friesen have later developed the EMFACS system, which specifies which facial actions are common for particular emotion displays. Obviously, the projection of emotional display to a low-dimensional space (as in the arousal/valence system) is an oversimplification, just like its

classification by a few discrete categories (like Ekman's basic categories). Yet, as the state-of-the-art in autonomous categorization progresses, finer grained representations will become possible for use in computer systems.

Table 1 Action Units (AU) in the Facial Action Coding System

AU	Descriptor		AU	Descriptor		AU	Descriptor
1	Inner Brow Raiser		2	Outer Brow Raiser		4	Brow Lowerer
5	Upper Lid Raiser		6	Cheek Raiser		7	Lid Tightener
8	Lips Towards Each Other		9	Nose Wrinkler		10	Upper Lip Raiser
11	Nasolabial Deepener		12	Lip Corner Puller		13	Cheek Puffer
14	Dimpler		15	Lip Corner Depressor		16	Lower Lip Depressor
17	Chin Raiser		18	Lip Puckerer		19	Tongue Out
20	Lip Stretcher		21	Neck Tightener		22	Lip Funneler
23	Lip Tightener		24	Lip Pressor		25	Lips Part
26	Jaw Drop		27	Mouth Stretch		28	Lip Suck
29	Jaw Thrust		30	Jaw Sideways		31	Jaw Clencher
32	Lip Bite		33	Cheek Blow		34	Cheek Puff
35	Cheek Suck		36	Tongue Bulge		37	Lip Wipe
38	Nostril Dilator		39	Nostril Compressor		43	Eyes Closure
44	Squint		45	Blink		46	Wink
(40)	Eyes Normally Open		(41)	Lid Droop		(42)	Slit

There are several other concerns in categorizing expressions. The manifestation of affective states depends on a particular **context**. The specification of the context serves the dual purpose of disambiguation of the expressive display, and constraining the search space for the expression. For instance, the AmI project (Carletta, 2006) focuses on a particular meeting scenario, and the corpus that is collected within this application framework is annotated using the following categories: neutral, curious, amused, distracted, bored, confused, uncertain, surprised, frustrated, decisive, disbelief, dominant, defensive, supportive. These categories are not recognized as universal expressions, yet they are identifiable and consistently labelled within the particular meeting scenario context.

Expressions also have **temporal dimensions**. We do not feel surprise for a second, followed by a short burst of happiness, and immediately switch to disgust. A system that continuously evaluates affect needs to take into account the temporal unfolding of the expressions. An additional benefit of this approach would be to account for the "baseline effect", which stresses the relevance of the difference from the neutral state as opposed to the absolute feature locations for identifying emotions, especially when they are subtle. The differences and changes in the facial configuration can be used to describe emotional displays at a much higher granularity, allowing realistic contextual modelling. Going to contextual level is promising for two reasons: The actual switch from one expression to another can be more accurately recognized in the presence of other contextual cues. Conversely, reliable detection of an expression change can point to an external event of importance, thus leading to improved event categorization.

Most of the earlier research focuses on **posed** (or simulated) expressions. For better discrimination, the expressions under study are created by persons imitating a certain expression in an exaggerated way. There is however subtle differences between faked expressions and expressions caused by true emotions. Some recent research is tailored towards making this distinction explicit, for instance to understand whether an expressed emotion is honest or not. The temporal nature of emotional expressions also allows one to test for the authenticity of an emotional display. The so–called micro–expressions are involuntary

cues that are persistently found in genuine expressions, yet are usually absent from faked ones. However, the spatio–temporal resolution of these cues is very fine, and it is very difficult to separate them from measurement noise, as their magnitudes are comparable under currently used experimental settings. On the other hand, a fake expression involves more conscious control, and thus is not prone to correlate highly with naturally occurring bodily gesture cues. Subsequently, multimodal analysis becomes a promising alternative to detect genuine emotion expressions.

An important point here is that people are equally successful in recognizing the category of genuine and posed emotional expressions. This suggests that posed expressions, while not exactly overlapping with their genuine counterparts, have their own semiotic function that works within the appropriate social context. It follows that an all-encompassing emotion recognition software needs to model both genuine and posed expressions, and to recognize both.

INTERPRETATION OF AFFECT FROM FACIAL CUES IN HUMANS

In this section we look at some biological and cognitive aspects of facial expression recognition in humans. We should at this point stress that the subjective feeling of an emotion and its expression on the face are two different things, where the latter is one manifestation of the former among many bodily signals like gestures, postures, and changes on the skin response. Thus, what we perceive from a face is either an involuntary manifestation of an emotional state, or the result of a deliberate effort at communicating an emotional signal. The urge to associate affect with faces is so great that we 'recognize' expressions even on infant's faces, even though they are not yet associated with the emotions they represent in adults (See Figure 2). This association partly relies on innate biases implicit in the human visual system, and partly on the efficient way humans represent facial information. In humans, the subjective experience of an emotion, the production of its somatic expressions, and its recognition in other subjects are all tightly coupled, and influence each other. This allows for a degree of feedback that is beyond current computer systems, and enables differentiation of very subtle affective cues.

Figure 2. A human face is rarely perceived as completely neutral; even a two-weeks–old infant's face is full of "emotions".

The goal of facial affect recognition systems is to mimic humans in their evaluations of facial expression. If a computer can learn to distinguish expressions automatically, it becomes possible to create interfaces

that interpolate affective states from these expressions and use this information for better interfaces. We open a little parenthesis here. When we talk about 'learning' in the context of a computer, we usually mean a machine learning procedure, which is different from human learning. Here, what usually happens is that the computer is provided with a number of samples from a category to be learned (be it images of faces with a particular expression or any other numeric representation), as well as a method of categorization. The learning algorithm tunes the parameters of the method to ensure a good categorization on these samples. The ensuing system, however, depends crucially on the quality of provided samples, in addition to the data representation, the generalization power of the learning method and its robustness to noise and incorrect labels in the provided samples. These points are shared by all computer systems working on face images, be it for the recognition of identity or expressions. We bear these in mind when investigating what the brain does with faces, and how it can be simulated with computers.

Recognition of relevant processes that partake in human recognition of faces and facial affect guides the designers of computer algorithms for automatic recognition of emotions from faces. For instance, it is known that humans have selective attention for the eyes and mouth areas, which can be explained by recognizing the importance of these areas for communicating affect and identity. Computer simulations by Lyons et al. (1999) have shown that feature saliency for automatic algorithms that evaluate facial affect parallels feature saliency for the human visual system.

How humans determine identity from faces is a widely researched area. One reason for this is that both low-level neurological studies and high-level behavioural studies point out to faces as having special status among other object recognition tasks. Kanwisher et al., (1997) have argued that there is an innate mechanism to recognize faces, and they have isolated the lateral fusiform gyrus (also termed the fusiform face area) to be the seat of this process. The proponents of the expertise hypothesis, on the other hand, argued that humans process a lot of faces, and this is the sole reason that we end up with such a highly specialized system (Gauthier et al., 1999).

The expertise hypothesis banks on a fundamental property of the human brain: the key to learning is efficient representation, and while we learn to recognize faces, the neural representation of faces gradually changes, becoming tailored to the use of this information. In other words, we become (rather than born as) face experts. But this also means that we are sensitive to cultural particularities we are exposed to, an example of which is the famous other-race effect. This is also true for affect recognition from facial expressions, which incorporate cultural elements. While the geometric and structural properties of a face might allow the viewer to distinguish the basic emotional content, cross–cultural studies have established that the cultural background of the viewer plays a large role in labelling the emotion in a face. Furthermore, perception of emotion-specific information cued by facial images are also coloured by previous social experience. In a recent study (Pollak et al., 2009), a number of children who have experienced a high-level of parental anger expression were shown sequences of facial expressions. They were able to identify the anger expression in the sequence earlier than their peers, using a smaller amount of physiological cues.

The traditional problems faced by face recognition researchers are illumination differences, pose differences, scale and resolution differences, and expressions (See Figure 3). These variables change the appearance of the face, and make the task of comparing faces non-trivial for the computer. While there is a consensus among brain researchers that recognizing facial identity and facial affect involve different brain structures (e.g. lateral fusiform gyrus for identity as opposed to superior temporal sulcus for emotional content, (Hasselmo, Rolls & Baylis, 1989)), these are not entirely independent (Bruce & Young, 1986). Many aspects of facial identity recognition and affect recognition overlap. This is also the case for computer algorithms that are created for recognition of identity or affect from face images. Hence, it should be no surprise that computational studies also recognize the need for different, but overlapping representations for these two tasks. For instance Calder and colleagues (Calder et al., 2001)

have investigated a popular projection based method for classifying facial expressions, and determined that the projection base selected to discriminate identity is very different than the base selected to discriminate expressions. Also, while facial identity concerns mostly static and structural properties of faces, dynamic aspects are found to be more relevant for emotion analysis. In particular, the exact timing of various parts of an emotional display is shown to be an important cue in distinguishing real and imitation expressions (Cohn & Schmidt, 2004). Similarly, the dichotomy of feature–based processing (i.e. processing selected facial areas) versus holistic processing (i.e. considering the face in its entirety) is of importance. Features seem to be more important for expressions, and while in some case it can be shown that some expressions can be reliably determined by looking at a part of the face only (Nusseck et al., 2008), the dynamics of features and their relative coding (i.e. the holistic aspect) cannot be neglected.

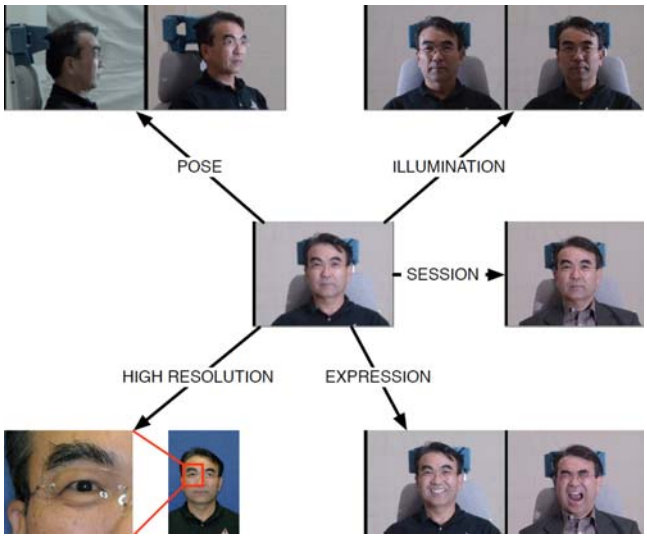


Figure 3. Variations captured by the state-of-the-art Multi-PIE database. © 2010, Gross et al. Used with permission.

Before moving to tools and techniques for computer analysis of facial expressions, we note here that all emotions were not created equal. Brain studies suggest different coding mechanisms for particular emotions. According to the valence hypothesis there is a disparity between the processing of positive and negative emotions, as well as the amount of processing involved for these types in the left and right hemisphere of the brain (Borod et al., 1998). This is an evolutionarily plausible scenario, as rapid motor response following particular emotions (e.g. fear, anger) is important for survival. Blair et al. (1999) have found that the prefrontal cortex is more active for processing 'anger', as opposed to 'sadness'. Different cortical structures show differential activation for different emotion types under lesion and functional imaging studies. On the other hand, specific emotions do share common neural circuitry, as disproportionate impairment in recognizing a particular emotion is very rare, as shown by lesion studies (the reader is referred to (Adolphs, 2002) for examples and references).

This inequality is also reflected in displays of emotion. The configural distances from a neutral face are disproportionate for each emotion, with 'sadness' and 'disgust' being represented by more subtle changes (as opposed to for instance 'happiness' and 'fear'). In addition to this disparity, it is unlikely that emotions are encountered with the same background probability in everyday life. Thus, from a probabilistic point of view, it makes sense not to treat all six basic emotions on the same ground. The valence hypothesis suggests that 'happiness' (as a positive emotion) is a superordinate category, and should be pitted against negative emotions (fear, anger, disgust, sadness and contempt). Surprise can be divided into 'fearful surprise' and 'pleasant surprise'; it has been noted that 'surprise' and 'fear' are often

confused in the absence of such distinction. Also, 'disgust' encompasses responses to a large range of socially undesirable stimuli. When it expresses disapproval for other people, for instance, it approaches 'anger'. These issues require careful attention in the design and evaluation of computer systems for facial expression analysis.

A STARTER'S KIT FOR COMPUTER ANALYSIS OF FACIAL EXPRESSIONS

In this section we give pointers to relevant methods for facial expression analysis in the literature and summarize the most important techniques, as well as challenges. Evaluation of facial expression relies on accurate face detection, face registration, localization of fiducial points in faces, and classification of shape and appearance information into expressions. As such, recognizing emotional expressions from faces can be treated as a pattern recognition problem.

For a very extensive list of different approaches (including visual, as well as audio-based and multimodal approaches) to affect recognition, the reader is referred to (Zeng et al., 2009). For a more focused and very readable survey of facial expression analysis, see (Fasel & Luettin, 2003). Figure 5 summarizes the information flow of a facial expression analysis system, where each stage is annotated with design decisions and difficulties from this perspective (dashed boxes). The expressive face image contains variations due affective state, as well as variations due pose, illumination, scale and resolution. The data acquisition itself adds some noise, which can be significant in natural settings. Face detection is generally the first step in the processing pipeline, followed by determination of the pose, either via localization of facial landmarks, or via iterative fitting of a face model to the appearance. The analysis of the face image needs to dissociate subject-specific variation from expression-induced changes. Analysis of features (static or dynamic) can be supplemented with context, and via information fusion with other modalities.

Face detection

Face detection is a crucial first step in facial expression analysis (Yang et al., 2002). The state-of-the-art in face detection is the Viola-Jones algorithm (Viola & Jones, 2004) that is freely available in the OpenCV library (see Databases and Tools section). The key idea behind this algorithm is to use a hierarchical Adaboost cascade classifier, which eliminates locations that obviously do not contain face images quickly, and to focus on likely candidates. A multi-resolution Haar wavelet basis is used, which is simple and fast to compute. The training is accomplished on a very large number of positive and negative samples (about forty thousand images for the OpenCV cascade).

The Viola-Jones algorithm has its limitations; since it is essentially a 2D face detector, it can only generalize within the pose limits of its training set. Also, large occlusions will impair its accuracy. It is however possible to train other cascades for faces with different pose variations with this method, as well as cascades for individual facial features. The one big advantage of this algorithm is in its speed, which is essential for real-time expression analysis. Additional cascades and more accurate detectors will come at a computational cost. (Bartlett et al., 2005) presents some improvements to this algorithm, for which the code is also freely available.

Facial feature localization

Analysis of detected faces often proceeds by locating several fiducial points on them. These features are called anchor points, or landmarks. Typically, eye and eyebrow corners, centre of iris, nose tip, mouth corners, and tip of the chin are located for face alignment. For expression analysis, a greater number of landmarks are usually required (typically between 20-60). Collectively, the landmark locations define the *shape* of the face, whereas the texture of the face is called its *appearance*.

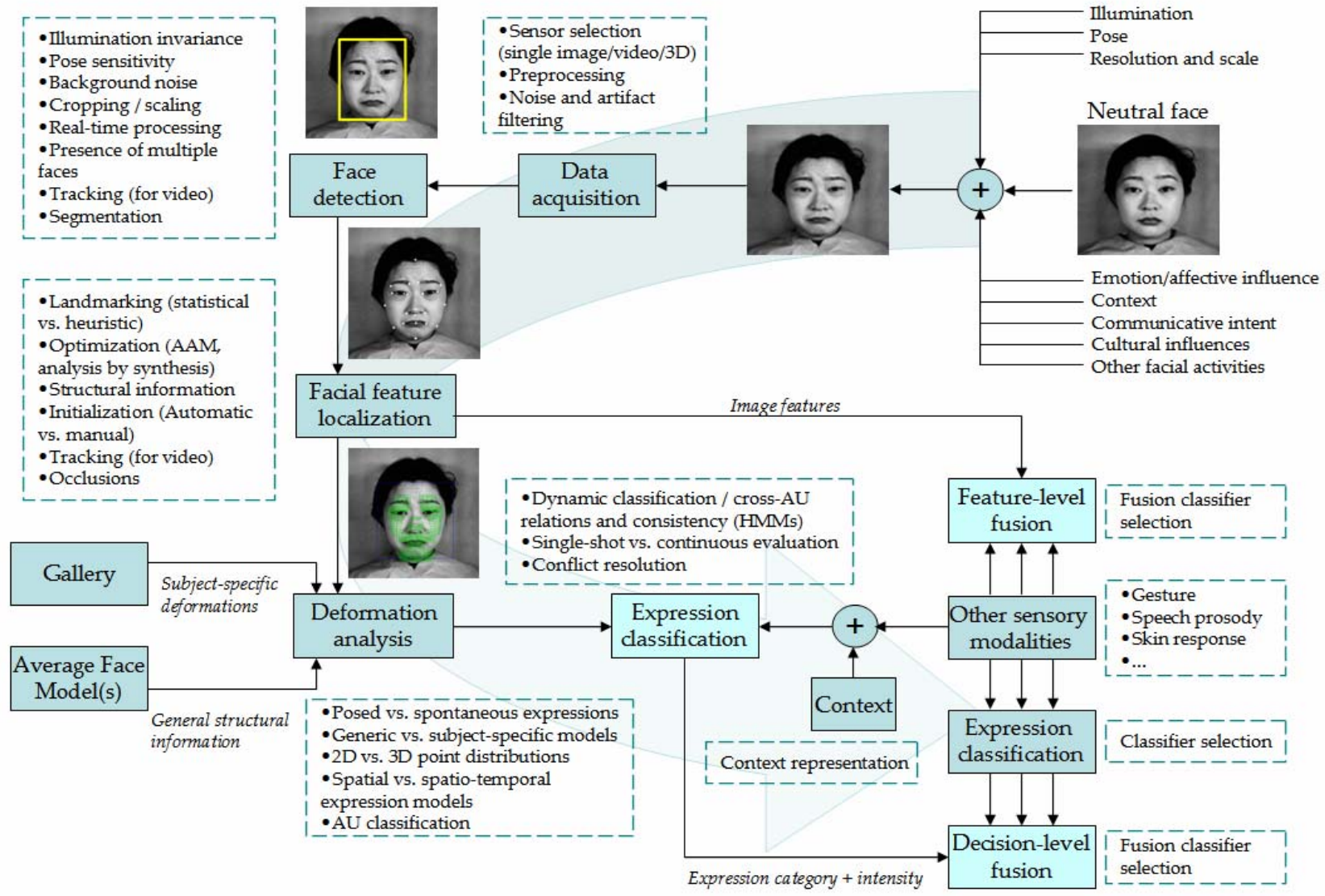


Figure 4 Overview of computer analysis of facial expressions. Face images taken from the Jaffe database.

The configurations of facial landmarks are indicative of deformations caused by expressions. Subsequently, deformation analysis can reveal expression categories, provided that facial landmarks are accurately detected and they contain sufficient information for the recognition of a particular facial expression.

Finding facial landmarks automatically is a difficult problem, which faces all hurdles of face recognition in a smaller scale, like illumination and occlusion problems (Salah et al., 2007). Constellation of facial landmarks is different for each face image. Part of the difference is due to the subjective morphology of the face, as different persons have different face shapes. Even for the same person, different images will have different configurations. Another part of this difference is due to camera angle and pose differences. There are also expression-based changes (of which some part may be attributable to emotion) and measurement noise, which is omnipresent. Commercial facial landmarking applications rely on tens of thousands of manually annotated images for training robust and fast classifiers.

The appearance of each landmark and the structural relationships between landmark points (i.e. configuration) are both taken into account in locating landmarks automatically. However, both appearance and structure is changed under expression variations, and in different ways. For this reason, most methods solve the simpler problem of landmarking the neutral face, and then track each landmark while the face is deformed under the influence of an expression. The deformation sequence then allows one to classify the expression category.

While landmark-based approach is the mainstream in 2D facial expression analysis, a recent study has shown that facial landmark distributions are not always sufficient to distinguish emotional expressions (Afzal et al., 2009). In this study, the authors asked their subjects to label faces according to five emotional expressions (interested, bored, confused, happy, and surprised). The contrasted representations were 1) face videos, 2) a point-light representation of the facial landmarks, 3) a stick-figure of these landmarks (appropriately connected), and 4) the mapped expression display with an Xface virtual agent (See Figure 5). Their findings show that 1) natural expressions are harder than posed expressions to identify, 2) expressions are best identified in the original videos, followed by stick-figure, point-light and finally virtual agent representations, 3) subjective labels of expression categories are moderate for the original videos, but quite poor for other representations, 4) the accuracies vary greatly across different expressions.

Figure 5. The original frame and three synthesized representations. © 2009, Afzal et al.Used with permission.

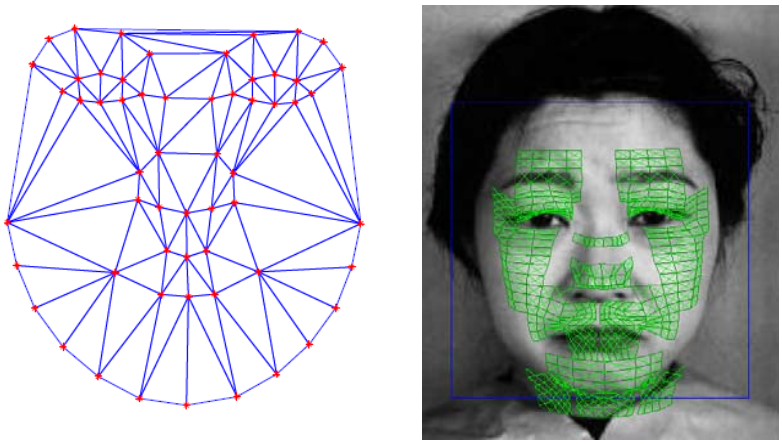
Expression classification from images and video

A large number of detection and classification approaches are developed for categorization of facial expressions from images. Typically, the first steps are face detection, landmark localization, and pre-processing for scale and illumination normalization. For image-based methods, if the neutral configuration of the face is not known, the appearance of the face (either global appearance, or features extracted from the landmark points), as well as configuration of the shape relative to an average shape are

informative cues to classify expression. Different parts of the face contain different amounts of information, making segmentation into different regions a reasonable choice. For instance Figure 6 (b) shows a possible segmentation of the face based on relevance to expression recognition, as well as to facial motion.

A popular approach for face analysis that relates shape and appearance is the active appearance model (AAM) (Cootes et al., 2001). In this approach, the 'shape' (represented by a set of connected landmarks, see Figure 6 (a)) and the 'appearance' of a face are jointly modelled. AAM is essentially an analysis-by-synthesis method; the correct parameterisation that represents a novel face is obtained by synthesizing face images from a generative model until these images look sufficiently close to the analysed image.

The method requires the initial generation of a shape and an appearance model. For this purpose, a set of training images are used, on which several landmarks are annotated. These images are rigidly aligned to an averaged shape, which is just a set of landmark locations. Then, each training image is warped to this shape. The 'appearance' model is obtained separately from the texture. In this method, a novel face can be represented by a mean shape and a mean appearance, plus linear combinations of shape and appearances present in the training set.



(a) (b)

Figure 6. (a) Landmarks and corresponding triangular mesh model for AAM (Stegmann, 2002). (b) The wireframe model used in (Cohen et al., 2003).

In practice, when a new face image is analysed with the AAM, the search for the best parametric representation is initialized with an average face. Then, the residual error is minimized by adjusting the parameters of the generative model iteratively, in a coarse-to-fine fashion, each time synthesizing a new face with modified shape and appearance parameters. When the best parameter configuration is obtained, the shape parameters indicate the landmark locations. These can then be used in classifying the expression of the face. Since AAM is not a full-blown 3D model, its accuracy starts to diminish rapidly for faces with pose differences beyond ±20º from the frontal pose. A detailed study of appearance models and subsequent modifications can be found in (Lee et al., 2009). Recently, Milborrow and Nicolls (2008) extended the AAM to create a very successful system for automatic facial landmark localization on frontal faces. This system requires over 60 annotated landmarks for each face during training, but once trained, shows remarkable generalization for frontal and neutral faces acquired in different conditions (see Table 3).

A related methodology is the analysis-by-synthesis approach, which tries to synthesise a face image that matches a query image as closely as possible, and optimizes the pose, illumination and expression parameters of a generative model for this purpose (Volker & Blanz, 1999). These unknown parameters

are estimated iteratively, and a fully 3D model is maintained, which make this approach computationally expensive. There is a vast literature in face and facial expression synthesis, which are excluded from the present survey.

While earlier approaches favoured holistic analysis of the face images, AAM and other model-based approaches eventually received more attention. Another trend that exhibits itself is the extensive use of neural network classifiers in earlier work, as opposed to support vector machines (SVM), AdaBoost and dynamic Bayesian network (DBN) classifiers in more recent approaches. For instance (Bartlett et al., 2005) compares AdaBoost, SVM and linear discriminant analysis (LDA) classifiers on a range of features. The best results are obtained by initially deriving a very large number of features combined with clever feature selection methods. In their study, the authors use a bank of Gabor wavelet filters at 8 orientations and 9 spatial frequencies on face images scaled to 48x48 pixels. This processing results in about 160,000 features per face. Then the AdaBoost algorithm is used as a feature selection method to choose 900 features, which are fed to an SVM classifier. A number of AU-detectors implemented with this approach are combined in a second stage to detect driver drowsiness (Vural et al., 2007) and to distinguish between real and faked pain (Littlewort et al., 2009). In (Whitehill et al., 2009), a similar setup is used for a smile detection application, but Haar wavelet filters (also called Box filters) are contrasted with Gabor filters. The authors remark that an AU-detector needs to be trained with 1,000-10,000 training samples for robust operation.

The dynamic nature of expressions results in improved detection from video, where the spatio–temporal dynamics of facial structures can be evaluated. The Cohn Kanade database of FACS annotated video sequences of emotions has been a major facilitator of facial expression research from dynamic cues (Kanade et al., 2000). Major tools for implementing such systems include algorithms for motion flow field computation and tracking and Bayesian approaches, for instance Markov models for characterizing dynamics of emotional states. In optical flow methods, the spatio-temporal motion-energy templates are taken as characteristic for expression classes, and used for recognition (Essa & Pentland, 1997). Kalman filters and realistic muscle models can be used to increase the stability of the optical flow model. (Fasel & Luettin, 2003), as well as (Cohen et al., 2003) include good surveys of earlier dynamic approaches to facial expression recognition. Here we describe a few relevant methods.

In (Zhang & Ji, 2005), facial features are tracked with the help of infrared cameras, and a number of muscle-movement-based heuristics. Following (Tian et al., 2001), a Canny edge detector is used to enhance the furrows of the face. A number of rules are employed to infer AUs from the tracked feature points. The AU detection results are combined with a DBN. The inference performed by the DBN takes into account the dynamic nature of AU transitions, and improves the final detection rate. In a subsequent work, (Tong et al., 2007) combine the Gabor-AdaBoost feature selection with DBNs. They select 200 Gabor wavelet coefficients from the processed face images via AdaBoost, but then discretize the continuous output of the AdaBoost classifiers into binary values. If the value of a particular classifier is 0, it means that particular AU is not present, where a value of 1 indicates the presence of the AU. These binary values are used as evidence in the DBN learned from the training set.

In (Valstar & Pantic, 2006), the Gabor-AdaBoost scheme is supplemented with a particle filter based tracker. This method robustly tracks the feature points over subsequent frames, albeit at a higher computational cost. In (Koelstra & Pantic, 2008) features are detected in the first frame, and tracked over the subsequent frames, where classification is performed with Gentleboost (a variant of AdaBoost) and HMMs in succession.

In (Cohen et al., 2003), a model-based face tracking approach is taken, where a wireframe model of the face is constructed (see Figure 6 (b)). In the first frame of the image sequence, facial features such as the eye corners and mouth corners are selected by the user. Assuming a neutral, frontal pose, this part of the

model can also be automatically performed. The generic face model is then warped to fit the located facial features. Once the model is constructed and fitted, head motion and local deformations of the facial features such as the eyebrows, eyelids, and mouth can be tracked. The motion direction and intensity of the tracked points are used as observation vectors of a hidden Markov model (HMM). Each expression type has its own Markov model, and the model that produces the highest posterior probability for a given observed sequence is selected as the expression class. The authors also propose a second level HMM to model transitions between emotions, and use this model for automatic segmentation of the video sequences.

Expression classification from 3D faces

With the advances in 3D sensor technology 3D has become an attractive modality for face recognition. State–or–the–art databases created for 3D face recognition pay attention to expressions and facial movement. Figure 7 shows samples from the Bosphorus 3D face database, which is publicly available (see Databases and Tools section).

It is possible to process 3D faces in a way similar to 2D faces. A depth map representation can be obtained by mapping the depth of each 3D facial surface point to an intensity value, depending on the distance from the camera. This is called a 2.5D representation, and most 2D analysis techniques can be adapted for this representation. In order to harness the full power of the 3D representation, however, the points that represent the facial surface are aligned to a prototypical shape. While 3D information allows the computation of features like curvatures and shape indices, accurate landmark localization and deformation analysis remain to be challenging.

Figure 7. Action units and basic expressions from the Bosphorus 3D face dataset. © 2008, Alyüz et al. Used with permission.

There are relatively few purely 3D approaches for expression analysis in the literature. (Soyel & Demirel, 2007) use a neural network classifier that receives six indicative feature distances as an input (i.e. distances between eye and mouth corners, mouth and eyebrow height, and the face span) and categorizes the facial expression into a basic emotion category or as neutral. Subject-specific information is not used, and accurate landmarks (84 points) are assumed. Under these conditions, they achieve more than 90% classification accuracy. In (Mpiperis et al., 2008) 3D point clouds are aligned to a base mesh, but the mouth boundaries are detected separately for expressions with open mouth, as this causes a great change in the appearance. The classification is obtained by probabilistically modelling each expression class as a generative model. (Tang & Huang, 2008) use multi-class AdaBoost schemes with different weak classifiers, and obtain 95% average accuracy on the BU-3DFE database (Yin et al., 2006). This approach also relies on manually located landmarks.

It is obvious that there is room for improvement in 3D facial expression analysis. New scanner technologies also allow the analysis of 3D information with temporal dynamics (so–called 4D approaches) by enabling recording of 3D scans at rates approaching video acquisition (Yin et al., 2008). Recent work shows that processing dynamic information is also useful in the 3D modality (Sun & Yin, 2009). Consequently, it becomes possible to evaluate spontaneous expressions, which was not possible (or very difficult) for static 3D snapshots, acquired from highly controlled 3D scanner setups.

Databases and tools

Extensive research into face recognition has resulted in the collection of large numbers of datasets. Although most of these focus on the biometrics aspects, and are more suitable for identification experiments, some include facial expressions, and some are specifically tailored towards expression research. In Table 2 we summarize 12 databases that are open for research purposes. The primary modalities are single-shot images and videos depicting emotions. The Bosphorus database is obtained with a 3D scanner, and it is the most comprehensive 3D face database for expression research. The Canal 9 database is a collection of political debates, and provides for analysis in natural, as opposed to posed data collection settings.

Table 3 lists a number of open source (or freely distributed) software tools that are useful for facial expression research. Among these, we list tools for face detection and tracking (OpenCV, MPT), automatic facial feature point localization (Gabor-ffpd, STASM), active appearance models (AAM-API), expression annotation and labelling (SCORE, FEELTRACE), machine learning and pattern recognition (MPT, PRTools, WEKA, Torch, MSBNx). There are also links to embodied conversational agents (GRETA and XFace), and two repositories of tools and data from European framework projects (HUMAINE, SIMILAR/eNTERFACE).

Table 2 Databases for facial expression related research.

Database	Type	Details	Location
Cohn-Kanade (CMU-Pitt)	video, annot.	100 subj., neutral+ six basic emotions, 500 sequences, FACS action units	http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html
Green persuasive	video	eight discussions, 25-48 min. each, persuasiveness annotations	http://green-persuasive-db.sspnet.eu/
RU-FACS-1	video	100 subj., 2,5 min. recordings, FACS codes for 20% subjects, lie detection	http://mplab.ucsd.edu/?page_id=80
MPLab GENKI-4K	image	4000 images, annotated as smiling/non-smiling	http://mplab.ucsd.edu/?page_id=398
CMU PIE	image	68 subj., 41,368 images, 4 expressions, 43 illuminations, 13 poses	http://www.ri.cmu.edu/projects/project_418.html
Multi-PIE	image	337 subj., 750,000+ images, 6 expressions, 15 poses	http://multipie.org
Stegmann	image, annot.	37 subj. frontal neutral faces, 58 landmarks + shape model ground truth	http://www.imm.dtu.dk/~aam/datasets/face_data.zip
CAS-PEAL (R1)	image	1040 subj., 30.900 images, 5 expressions, accessories, 15 lighting directions	http://www.jdl.ac.cn/peal/
JAFFE	image	10 female Japanese models, 213 images of 7 expressions, intensity ratings	http://www.kasrl.org/jaffe.html
MPI	video	246 sequences of face actions taken simultaneously from 6 cameras	http://vdb.kyb.tuebingen.mpg.de/
Bosphorus	3d scans, image	105 subj., 4666 scans, 24 landmarks, action units and basic expressions	http://bosphorus.ee.boun.edu.tr/
BU-3DFE/ BU-4DFE	3d scans/ 3d video	100 subj., 2500 scans, basic expressions in four intensities, no landmarks 101 subj., 606 sequences of 100 frames each, basic expressions	http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html
MMI face	image, video	86 subj., 2894 sequences, posed and spontaneous expressions, audio incl.	http://www.mmifacedb.com/
Canal 9	video	72 political debates, total of 48 hours. speaker segmentation and group indication	http://canal9-db.sspnet.eu/

Table 3 Some software tools usable for facial expression related research.

Tool Name	Details	Location
OpenCV	C/C++ & Python library for real time computer vision (face detection)	http://sourceforge.net/projects/opencv/
MPT	C & MATLAB toolbox for machine perception tools (inc. eye detection and face tracking in video)	http://mplab.ucsd.edu/grants/project1/free-software/MPTWebSite/introduction.html
STASM	C++ library for finding features in frontal & neutral faces	http://www.milbo.users.sonic.net/stasm/index.html
Gabor-ffpd	MATLAB Gabor wavelet based facial feature point detector	http://www.doc.ic.ac.uk/~mvalstar/programs/sliwiga_ffpd.zip
AAM-API	C++ API for Active Appearance Models, related MATLAB tools	http://www2.imm.dtu.dk/~aam/
SCORE	Digital video coding and annotation tool inc. FACS annotations	http://mpscore.sourceforge.net/facs.php
FEELTRACE	Two dimensional (activation/evaluation) emotion labeling tool	http://emotion-research.net/download/Feeltrace%20Package.zip
Greta	Expressive embodied conversational agent, with face expression synthesis	http://perso.telecom-paristech.fr/~pelachau/Greta/
XFace	Expressive conversational agent, without a body	http://xface.itc.it/index.htm
PRtools	MATLAB toolbox for pattern recognition	http://www.prtools.org/
WEKA	Java toolbox of machine learning algorithms	http://www.cs.waikato.ac.nz/ml/weka/
Torch	C++ toolbox of machine learning algorithms	http://www.torch.ch/
MSBNx	Microsoft Bayesian Network editor and toolkit	http://research.microsoft.com/en-us/um/redmond/groups/adapt/msbnx/
HUMAINE	Source codes and data from HUMAINE network	http://emotion-research.net/toolbox/
eNTERFACE	Source codes and data from eNTERFACE Workshop series	http://www.enterface.net/results/

Applications and future trends

Widespread use of face detection and recognition technology has enabled a range of applications, some foreseeable in near future, and some desired applications with long–term research aspects. There are practical applications for recognizing each emotion separately; for instance recognition of fear and happiness has different implications.

In this section we shortly look at applications of facial expression analysis in several interrelated domains, including social analysis, robotics, ambient intelligence, and gaming. The boundaries of these applications are fuzzy; there are games for ambient intelligence settings, and robots for social analysis. Our aim here is to make the reader aware of some of the possibilities where this technology can take us in the near future. A comprehensive survey is beyond our scope.

Applications in social analysis

For social sciences that analyse human behaviours, affect is a relevant dimension that needs to be accounted for. While computers are not as accurate as humans in assessing affect, automatic recognition of emotion lightens the burden of costly and error–prone data annotation process, it enables analysis of datasets composed of long multimodal observations, and also stands to provide quantitative, objective measurements. With advances in automatic classification of FACS action units from videos, there have already been cases of computers outperforming humans. For example, in a recent study on pain (Littlewort et al., 2009), 170 naïve human subjects were shown videos of real and faked pain, and could differentiate between these classes only 49% of the time (with standard deviation 13.7%). The automatic system based on AU analysis on the other hand had 88% correct discrimination rate for the same task. Since pain is very subjective, it is not difficult to see that automatic tools to analyse pain from facial expressions would be very useful for diagnostic purposes. A similar application is assisting human training in distinguishing between real and faked expressions, which is not only a useful social skill, but a job requirement in some cases.

Imagine yourself being equipped with a device that can ‘read’ other’s facial expressions. Imagine further that you suffer from autism, and have trouble understanding expressions and interaction patterns of people around you. People with autism spectrum conditions (ASC) stand to gain much from such “empathy enhancing” technologies, especially if an unobtrusive and transparent interface can provide them with helping cues (El Kaliouby et al., 2006).

Clinical and psychological applications of facial expression analysis are not limited to autism research. (Ekman & Rosenberg, 2005) contains several studies relating to the analysis of depression, schizophrenia, psychosomatic disorders, suicidal tendencies, guilt expressions for psychotherapeutic interaction, and personality assessment. In these studies, facial expression is interpreted as a dependent variable of affect-related changes in the body, and automated tools are used in assisting diagnosis and therapy monitoring.

Affect assessment from facial cues can go beyond facial actions, as computers have access to sensors humans do not possess. For instance it has been shown that bio-heat modelling of facial imagery can reveal increased blood flow around the eyes (see Figure 8), which is a good indicator of stress, as well as cardiac pulse and heart rate (Pavlidis et al., 2007). These indicators can be used in clinical studies, in games that can adapt to the user’s stress levels, or even in criminal investigations.

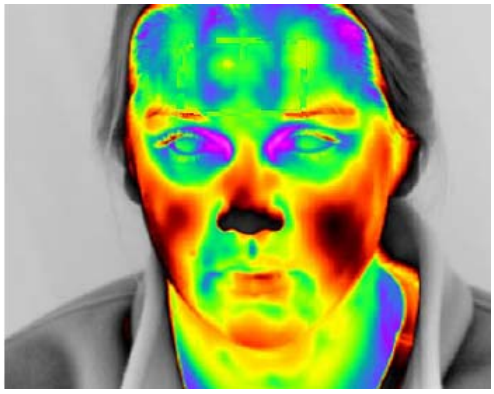


Figure 8. Thermal imaging of the face can reveal stress. © 2007, Pavlidis et al., adapted with permission.

Recent research directions in social signal processing employ facial expression analysis in assessment of group interaction dynamics. Relevant applications include automatic analysis of mood, coordinated patterns of interaction, mimicry, engagement, focus of attention, and dominance relations. Such indicators help for instance in automatic assessment of political discussions and campaign footages to predict the outcome of political debates, or for determining cognitive styles and personality aspects of individuals.

A related application is the automated analysis of impact for commercials. The cost of screening a commercial in a valuable slot (for instance during Super Bowl, the main sports event in US) can cost millions of dollars. Emotion–sensing technology has been harnessed for gauging the immediate impact of such expensive advertisements. It is also possible to measure affect directly in relation to the product. In one such application, Unilever has used video footages of people tasting different kind of food, and used the eMotion face expression analysis system developed at the University of Amsterdam (http://www.visual-recognition.nl/index.html) for obtaining objective and reproducible results.

A final application we would like to mention under this category is affect-based multimedia retrieval. Multimedia content analysis (MCA) tools enhanced with expression analysis can provide the means for qualitative search of material (querying for 'happy' episodes), highlight extraction, automatic life-logging and summarization (storing and retrieval of emotionally loaded content), and surveillance (retrieval of frames that contain people with angry or stressed expressions from a surveillance footage). The exponential growth of multimedia material accumulating on the Internet makes this type of affect-based indexing a very promising application (Hanjalic & Xu, 2005).

Applications in robotics

One of the goals of robotics is to create robots that interact naturally with humans. Understanding affect is a very important requirement for these applications. To give one example, consider robots designed to help autistic children, as partners of interaction or as educational tools to teach children basic visual concepts (Dautenhahn & Werry, 2004). (Salter, 2009) reports a case of robot–child interaction where the child is distressed by the music played by the robot, and signals this emotion through facial and bodily gestures. Unless the robot understands this signal, and acts upon it by terminating the activity that is causing the stress, the interaction will have undesired consequences.

Affect is an integral part of human communication, and sensitivity to affect allows for more natural interaction. For this reason, robotics researchers seek to endow robots with the ability to respond to human psychological states like fear, panic, stress, or to allow the robot build a more accurate representation of the interacting human by taking into account focus of attention, gaze direction, engagement, boredom and such properties. A more responsive robot presents a richer experience for the interacting party.

The dyadic nature of affective communication requires the evaluation of emotion, as well as an internal emotion model. A good example is MIT's Kismet robot, which has mechanisms for recognizing precipitating events, appraising it for affective content, displaying a certain expression through its face, voice and posture, and finally a set of action tendencies that motivate its behavioural response (Breazeal, 2001). Modulation of action selection mechanisms is particularly important, because an emotional display by the human needs to be acknowledged by the robot for seamless interaction. This will be done by modulating the behaviour of the robot appropriately.

Applications in ambient intelligence

Ambient intelligence (AmI) represents a vision where people are surrounded by smart appliances and devices that respond to overt and covert signals of the user in intelligent ways. It deals with both understanding of a user's emotions, and with synthesizing emotions on virtual agents to create the impression of affect for a more natural communication interface. As such, there are many AmI applications that can benefit from facial expression analysis. In ambient environments like **smart homes**, the recognition of affect improves categorization of action context. By correlating affect and intentions, it becomes possible to constrain the search for the correct interpretation of signals.

Detection of anger, fatigue or boredom of the driver in a **smart car** is desirable for increasing safety (Ji et al., 2006). A simple camera positioned behind the wheel allows tracking of the driver's face and analysing the expression for such signals. An assumption that gained much experimental evidence is that human errors are correlated with negative affective states. Consequently, detecting these states is a path to minimizing such errors.

Improving user's performance is a goal for many ambient intelligence technologies, including **personal wellness** and **assistive technologies**. Ambient intelligence settings are also adequate for **cognitive enhancement** applications, providing their users with useful information. One problem AmI needs to deal with, which was apparent even in its earliest applications, is the nuisance factor. People using such systems were getting annoyed at the 'smart' decisions taken by the system, which were sometimes badly timed. Conventional homes are predictable; when appliances around you start getting ideas, they may become unpredictable, causing frustration. Recognition of frustration in the user is one valuable skill for AmI applications.

Other AmI applications include **virtual guide systems**, and **virtual teachers**. The best tutoring systems have extensive feedback for the student, and they should be able to support, explain, evaluate, motivate, and provide expectations. Also, the social aspect of learning cannot be neglected; social interaction is a powerful catalyst for learning and needs to be harnessed for exploring new ways of teaching (Meltzoff et al., 2009). Autonomous systems that can provide such feedback can be the key to a revolution in education, making high-quality tutoring available to millions of people worldwide through computers. **Embodied conversational agents** (ECA) is an active research area for such virtual tutoring and guiding systems (Cassell et al., 2002, Ruttkay & Pelachaud, 2004). The aim is to create a virtual agent that is expressive enough to communicate appropriate affective cues to the user, thereby ensuring an improved communication experience. Figure 9 shows two such agents, XFace (Balcı, 2004) and GRETA (Ruttkay & Pelachaud, 2004), respectively.

Figure 9. Synthesizing affect on virtual head models Alice (Xface) and Greta. Alice displays anger, while Greta shows happiness here. See text for references.

Applications in gaming

Future gaming applications will have more input from the user, through increased sensor capabilities in end–user devices. Two types of developments in gaming are relevant through facial expression research, based on analysis and synthesis of expressions, respectively.

The first type of systems will try to understand the users affect through sensors on the computer or the game console, and subsequently adapt their behaviour appropriately. These systems will obviously require real-time processing, which can be challenging. They may have the advantage, however, of adapting to a specific user, which would mean lightweight feature processing. It is also conceivable that the user spends some time calibrating such a system, tuning it to his or her needs.

Multi-user online games are good examples for this category of game applications. The detection of affect and transfer thereof to the virtual agent (or avatar) controlled by the user is desirable for multi–user games in which multiple human players are embodied. Popular examples like World of Warcraft and Second Life boast millions of users.

With current technology, it is possible to have an avatar projecting real facial expressions in Second Life. The eMotion software mentioned earlier runs an expression recognition tool on the client side, and using the user interface of the program, sends automatic facial texture updates to the avatar. These updates correspond to read expressions of the user. Since the Second Life interface is not yet optimized for such input, the link between eMotion and Second Life is not through a clear interface. We may assume that in the future, multi-user games will implement the necessary stubs on the program side to receive affect from the client, and make it partially available to their virtual characters. Thus, it will be possible to register for example a ‘disgust’ reaction of the real user, captured on the client side by the affective system, recognize and encode this as ‘disgust’, and enact it on the avatar simultaneously.

The facial expressions can also serve as novel input modalities to a computer, allowing different gaming experience. The eMotion webpage mentioned before also includes a popular demo (also installed at the NEMO Science Museum in Amsterdam) where you can play a game of pong using facial expressions.

The second group of systems try to incorporate believable non-player characters, with a wide range of expressions. The game-playing experience is enriched by these emotional displays with possibly different intensities, with clearly visible or subtle signs, and with different frequencies of occurrence. The FACS

system, for instance, is already used in the game industry to synthesize realistic face expressions (e.g. in the Half-Life 2 game by Valve).

Naturally, there will be systems that combine both aspects. If a camera can be used to register the users disgust on the client side, and that information is conveyed to nearby virtual characters that are participating in the current interaction, these characters can act accordingly and modify their behaviours. For a more realistic gaming environment, it is also useful to have automatic characters that are able to respond to situations with pre–programmed or even learned semantics, and show affect in their facial and bodily expression.

CONCLUSIONS

There are several limitations of existing approaches to facial expression analysis. The evaluation is often conducted on posed expressions, with a small number of 'basic' emotions considered. Recent work in this area seeks to remedy this by considering natural data. However, manual annotation tools are not sufficiently developed to make annotation fast and cheap. Furthermore, if the temporal dimension and simultaneously recorded multimodal information are taken into account, it becomes apparent that annotation becomes much more difficult. Yet annotation is crucial for training statistical algorithms, as well as for evaluating methods against the golden standard of human judgement. Finally, recording high-resolution faces in isolation helps face detection and facial feature tracking, but it also means that contextual cues are neglected to a large part.

The current approach in social signal processing is to evaluate **dynamic** facial information in **natural contexts**. The results of such an approach will eventually influence psychological and neurological research, which predominantly work on static facial expressions to date. A future challenge is to create the tools for annotation and evaluation of dynamic and more granular affective content for psychological and neurological research. Also, it has been made clear that the transmission of human affect is a composite somatory event, and multimodal analysis of affect has much greater potential than unimodal analysis.

The communication of human affect does not solely rely on facial expression, but also on physical appearance, gestures, postures, spatio–temporal dynamics of behaviour, and vocal behaviour. **Multimodal analysis** takes into account these modalities, and stresses the importance and integration of contextual information (Sebe et al., 2005). For multimodal fusion with audio modality, the *prosody* is the singularly most used feature to complement visual information for expression analysis (Caridakis et al., 2010). (Zeng et al., 2009) includes a thorough survey of audio-visual methods for affect recognition. Other cues for fusion, while dependent on cultural context, include hand gestures (Yang & Ahuja, 2001), head motion (Cohn et al., 2004) and shoulder gestures (Valstar et al., 2007).

A machine learning system is only as good as its data; if the annotation is erroneous, the learning system will model incorrect correlations. Robust systems that can tolerate a certain level of noise in the data require increasing amounts of training data. **Collecting and publishing rich emotional data** is difficult, particularly as multimedia information would reveal personal and intimate information, and complex annotations are difficult and expensive to create. A potential solution is adapted by the SSPNet project (http://sspnet.eu), which aims at annotating and analysing existing videos of news and political debates. While alleviating the privacy issues, this approach has to deal with restricted context and imbalanced emotional content. Also, it foregoes the benefit of using auxiliary biosensors in automatically establishing ground truth for emotional content. However, the natural setting of these recordings and the relevance of the particular application makes it a worthwhile challenge.

One additional challenge is the **dissociation of facial affect** from speech–induced facial movement. This issue is rarely tackled, as most 'neutral' expressions in the available databases have closed mouths. Normal speech causes many deformations, which should not be recognized as emotional expressions.

Getting the **appropriate training data** is a challenge in many respects. Some affective states (like fatigue) can be induced, but some more complicated configurations are much more difficult to obtain. Humans can distinguish fine nuances of affective displays. A good example (given by Nick Campbell in a talk) is the following: "She was projecting happiness, but I could see she was unhappy". This kind of analysis is not surprising for us, it would be quite surprising if done by, say, a robot. Human-like understanding of affective states remains a grand challenge, especially since the optimal granularity of affect representation is not obvious. The categories we choose for computer classification can mimic linguistic levels, or they can be arbitrary groupings that we don't have words for.

Approaches motivated from a machine learning perspective are mostly interested in short-term correlations. However**, long-term within-subject correlations** (consistency of pain expression, for instance) are just as important as between-subject correlations, especially for clinical studies. This kind of analysis requires meticulous data collection and evaluation.

The bottom-up (or data-driven) approach can only take us so far in determining affect; we need top-down, **semantic information** to disambiguate patterns by also taking goals and environmental factors into account. The bottom-up approach essentially treats the problem as a classification task. Given a certain facial image, or a sequence of images, one (or multiple) classifications into affective categories are selected. While the sensory information contains physical, physiological, performance-related and behavioural cues, its relation to semantic indicators like goals, context or workload are not easy to assess, and the latter will have implications on the affective state of the individual. It may be particularly desirable to make inferences about the latter.

In spite of all these challenges, the availability of new tools, faster algorithms, more extensive databases, and the formation of research clusters and associations for affective computing (as well as new journals like IEEE Transactions in Affective Computing) make it a vibrant and rapidly progressing field.

ACKNOWLEDGEMENT

This work is supported by the EU NEST project PERCEPT.

REFERENCES

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms, *Behavioral and Cognitive Neuroscience Reviews*, 1(1), 21–62.

Alyüz, N., Gökberk, B., Dibeklioğlu, H., Savran, A., Salah, A. A., Akarun, L., & Sankur, B. (2008). 3D Face Recognition Benchmarks on the Bosphorus Database with Focus on Facial Expressions. *Proc. First European Workshop on Biometrics and Identity Management* (pp.62–71).

Afzal, S., Sezgin, T.M., Gao, Y. & P. Robinson. (2009). Perception of Emotional Expressions in Different Representations Using Facial Feature Points. *Proc. Int. Conf. Affective Computing and Intelligent Interaction*.

Balcı, K. (2004). Xface: MPEG-4 based open source toolkit for 3d facial animation. *Proc. Working Conference on Advanced Visual Interfaces*, (pp. 399-402).

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. Computer Vision and Pattern Recognition* (pp. 568–573).

Blair, R. J. R., Morris, J. S., Frith, C. D., Perrett, D. I., & Dolan, R. J. (1999). Dissociable neural responses to facial expressions of sadness and anger. *Brain*, 122, 883–893.

Blanz, V.& Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proc. SIGGRAPH,* (pp. 187–194).

Borod, J. C., Obler, L. K., Erhan, H. M., Grunwald, I. S., Cicero, B. A., Welkowitz, J., Santschi, C., Agosti, R. M., & Whalen, J. R. (1998). Right hemisphere emotional perception: Evidence across multiple channels. *Neuropsychology*, 12, 446–458.

Breazeal, C. (2001). Affective interaction between humans and robots. In Kelemen, J. & Sosík, P. (Eds.): *ECAL*, *Lecture Notes in Artificial Intelligence,* 2159, (pp. 582–591).

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology*, 77, 305–327.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W. & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, 41(9), 1179–1208.

Caridakis, G., Karpouzis, K., Wallace, M., Kessous L. & Amir, N. (2010). Multimodal user's affective state analysis in naturalistic interaction. *Journal on Multimodal User Interfaces*, 3(1), 49-66.

Carletta, J. (2006). Announcing the AMI Meeting Corpus. *The ELRA Newsletter,* 11(1), 3–5.

Cassell, J., Sullivan, J, Prevost S. & Churchill, E. (Eds.) (2002). *Embodied Conversational Agents*, Cambridge, MA: MIT Press.

Cohn, J., 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'l Conf. Systems, Man, and Cybernetics*, (pp. 610-616).

Cohn, J., & Schmidt, K. (2004). The Timing of Facial Motion in Posed and Spontaneous Smiles. *International Journal of Wavelets, Multiresolution & Information Processing*, *2*(2), 121–132.

Cohen, I., Sebe, N., Garg, A., Chen, L. S. & Huang, T. S. (2003). Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. *Computer Vision and Image Understanding*, 91(1-2), 160–187.

Cootes, T.F., Edwards, G.J. & Taylor, C.J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), 681–685.

Dautenhahn, K. & Werry, I. (2004). Towards interactive robots in autism therapy: background, motivation and challenges. *Pragmatics and Cognition*, 12(1), 1–35.

Ekman, P. (1993). Facial expression and emotion. *American Psychology* 48, 384–392.

Ekman, P. & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P. & Rosenberg, E. (Eds.) (2005). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Revised 2nd Edition. New York, NY: Oxford University Press.

El Kaliouby, R., Picard, R. & Baron-Cohen, S. (2006). Affective Computing and Autism. *Annals of the New York Academy of Sciences*, 1093(1), 228-248.

Fasel, B. & Luettin, J. (2003).Automatic facial expression analysis: Survey. *Pattern Recognition*, 36, 259–275.

Gauthier, I., Tarr, M. J., Aanderson, A., Skudlarski, P. & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, 2, 568–573.

Gross, R., Matthews, I., Cohn, J., Kanade, T. & Baker, S. (2010). Multi-PIE, *Image and Vision Computing,* 28(5), 807-813.

Hanjalic, A. & Xu, L.Q. (2005). Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1), 143-154.

Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face–selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research*, 32, 203–218.

Ji, Q., Lan, P. & Looney, C. (2006). A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on Systems, Man, and Cybernetics-A,* 36(5), 862-875.

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

Kanwisher, N., McDermott, J., & Chun, M.M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17, 4302-4311.

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*.

Lee, S.J. Park, K.R. & Kim, J. (2009). A comparative study of facial appearance modeling methods for active appearance models. *Pattern Recognition Letters*, 30(14), 1335-1346.

Littlewort, G. C., Bartlett, M. S. & Lee, K. (2009). Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*. 27(12), 1797-1803.

Lyons, M. J., Budynek, J. & Akamatsu, S. (1999), Automatic Classification of Single Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 21(12), 1357-1362.

Meltzoff, A., Kuhl, P., Movellan, J. R. & Sejnowski T. (2009). Foundations for a New Science of Learning. *Science*, 235(5938), 284–288.

Milborrow, S. & Nicolls, F. (2008). Locating Facial Features with an Extended Active Shape Model. *Proc. European Conference on Computer Vision*, 4, (pp. 504-513).

Mpiperis, I., Malassiotis, S. & Strintzis, M. G. (2008). Bilinear models for 3-D face and facial expression recognition. *IEEE Transactions on Information Forensics and Security*. 3(3), 498-511.

Nusseck, M., Cunningham, D.W., Wallraven, C. & Bülthoff, H.H. (2008). The contribution of different facial regions to the recognition of conversational expressions. *Journal of Vision*, 8(8), 1-23.

Ortony, A., Clore, G., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Pavlidis, I., Dowdall, J., Sun, N., Puri, C., Fei, J. & Garbey, M. (2007). Interacting with human physiology. *Computer Vision and Image Understanding*, 108(1-2), 150-170.

Pollak, S.D., Messner, M., Kistler, D. J. & Cohn, J. F. (2009). Development of perceptual expertise in emotion recognition. *Cognition*, 110(2), 242–247.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161-1178.

Russell, J. A. & Fernández–Dols, J. M. (Eds.). (1997). *The Psychology of Facial Expression*. Cambridge: UK, Cambridge University Press.

Ruttkay, Z. & Pelachaud, C. (Eds.) (2004). *From Brows till Trust: Evaluating Embodied Conversational Agents*, Kluwer.

Salah, A.A., Çınar, H., Akarun, L. & Sankur, B. (2007). Robust Facial Landmarking for Registration. *Annals of Telecommunications*, 62(1-2), 1608-1633.

Salter, T. (2009). A Need for Flexible Robotic Devices. *AMD Newsletter,* 6(1), 3.

Sebe, N., Cohen, I & Huang, T. S. (2005). Multimodal emotion recognition. In *Handbook of Pattern Recognition and Computer Vision,* World Scientific.

Stegmann, M.B. (2002). Analysis and segmentation of face images using point annotations and linear subspace techniques. Technical Report, Informatics and Mathematical Modelling, Technical University of Denmark (DTU). Retrieved 19 September 2009, from http://www2.imm.dtu.dk/~aam/.

Soyel, H. and Demirel, H. (2007). Facial expression recognition using 3D facial feature distances. *Lecture Notes in Computer Science*, 4633, 831-843.

Sun, Y. & Yin, L. (2009). Evaluation of spatio-temporal regional features for 3D face analysis. *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops.*

Tang, H. and Huang, T.S. (2008). 3D facial expression recognition based on automatically selected features. *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*.

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

Tong, Y., Liao, W. & Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10), 1683-1699.

Valstar, M. & Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. *Proc. Computer Vision and Pattern Recognition Workshop*.

Valstar, M.F., Güneş, H. & Pantic, M. (2007). How to Distinguish Posed from Spontaneous Smiles Using Geometric Features. *Proc. ACM International Conference Multimodal Interfaces*, (pp. 38-45).

Vinciarelli, A., Pantic, M. & Bourlard, H. (2009). Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing*, 27(12), 1743-1759.

Viola, P. & Jones, M.J. (2004). Robust real-time face detection. *International Journal of Computer Vision,* 57(2), 137-154.

Vural, E., Çetin, M., Erçil, A., Littlewort, G., Bartlett, M. & Movellan, J. (2007). Drowsy driver detection through facial movement analysis. *Lecture Notes in Computer Science*, 4796, 6-19.

Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M. & Movellan, J. (2009). Towards Practical Smile Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 31(11), 2106-2111.

Yang, M.H. & Ahuja, N. (2001). *Face Detection and Gesture Recognition for Human – Computer Interaction*. Kluwer Academic Publishers.

Yang, M.H., Kriegman, D., & Ahuja, N. (2002). Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1), 34–58.

Yin, L., Chen, X., Sun, Y., Worm, T. & Reale, M. (2008). A High-Resolution 3D Dynamic Facial Expression Database. *Proc. 8th Int. Conf. on Automatic Face and Gesture Recognition*.

Yin L., Wei X., Sun Y., Wang J. & Rosato, M. J. (2006). A 3D facial expression database for facial behavior research. *Proc. Int. Conf. on Automatic Face and Gesture Recognition,* (pp.211–216).

Zeng, Z., Pantic, M., Roisman, G. I. & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39–58.

Zhang, Y. & Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5), 699–714.

ADDITIONAL READING

The classification of facial expression has been investigated in several works:

- Ekman, P. & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial expressions*. CA: Consulting Psychologists Press.

- Izard, C. E. (1979). *The Maximally Discriminative Facial Movement Coding System (MAX)*. Newark, DE: University of Delaware, Instructional Resources Centre.
- Fridlund, A. J. (1994). *Human facial expression*. New York, NY: Academic Press.
- Ekman, P. (2003). *Emotions revealed*. New York: Times Books.

Theories of emotion and communication are linked to the study of facial affect:

- Mehrabian, A. (1972). *Nonverbal Communication.* Chicago, IL: Aldine Atherton.
- Whissell, C. (1989). *The dictionary of affect in language*. New York, NY: Academic Press.
- Baron–Cohen, S. Golan, O., Wheelwright, S. & Hill, J.J. (2004). *Mind reading: The interactive guide to emotion*. London: Jessica Kingsley Publishers Ltd.
- Scherer, K. R. (2000). Psychological models of emotion. In J. C. Borod (Ed.), *The neuropsychology of emotion* (pp. 137–162). New York, NY: Oxford University Press.
- Schmidt, K. L. & Cohn, J. F. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *Yearbook of Physical Anthropology*, 44, 3– 24.
- Coan, J. A. & Allen, J. J. B. (2007). *Handbook of Emotion Elicitation and Assessment*. New York, NY: Oxford University Press.
- Lewis, M., Haviland-Jones, J.M., Barrett, L.F. (Eds.) (2008). *Handbook of Emotions*. New York, NY: The Guildford Press.

A few very readable introductions give useful insights into the challenges and vision of automatic facial expression analysis:

- Cohn, J.F. (2006). Foundations of Human Computing: Facial Expression and Emotion. Presented in *Int. Conf. on Multimodal Interfaces*, (pp.233–238).
- Pantic, M. (2009). Machine Analysis of Facial Behaviour: Naturalistic and Dynamic Behaviour, *Philosophical Transactions of Royal Society B,* 364, 3505-3513.

Spatiotemporal and multimodal analysis is the future of human-computer interaction:

- Cowie, R., Douglas–Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J.G. (2001). Emotion Recognition in Human–Computer Interaction, *IEEE Signal Processing Magazine*, 18(1), 32–80.
- Pantic, M., Pentland, A. Nijholt, A. & Huang. T. (2008). Human–centred intelligent human–computer interaction (HCI2): How far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2), 168–187.

Instilling computers with emotion-related capabilities was frowned upon, until R. Picard made her case:

- Picard, R. W. (1997). *Affective Computing.* Cambridge, MA: MIT Press.
- Picard, R. W., Vyzas, E. & Healey, J. (2001). Toward machine emotional intelligence: analysis of affective physiological state, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 1175–1191.
- Picard, R.W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D. & Strohecker, C. (2004). Affective learning — a manifesto, *BT Technology Journal*, 22(4), 253–269.

Affective computing mostly focuses on individuals, whereas social signal processing emphasizes interaction and social dynamics. It adds social semantics to the processing of social signals:

- Vinciarelli, A., Pantic, M. & Bourlard, H. (2009). *Social Signal Processing: Survey of an Emerging Domain*. Journal of Image and Vision Computing.
- Thiran, J. –P., Bourlard, H. & Marques, F. (Eds.) (2009). *Multimodal Signal Processing*, Academic Press.

- Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12), 1775-1787.

Several surveys of facial expression recognition (in addition to the more recent (Zeng et al., 2009) cited in the main references) are worth mentioning:

- Pantic, M. & Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), 1424–1445.
- Tian, Y., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. In S. Z. Li & A. K. Jain (Eds.), *Handbook of face recognition*, pp. 247–276. NY: Springer.
- Güneş, H., Piccardi, M. & Pantic, M.. (2008). From the Lab to the Real World: Affect Recognition Using Multiple Cues and Modalities. In J. Or (Ed.) *Affective Computing: Focus on Emotion Expression, Synthesis and Recognition* (pp.185–218), Vienna: Austria, I–Tech Education and Publishing.

Recent encyclopaedic work in affective computing includes several comprehensive volumes. These are broader in scope, and include facial affect:

- Davidson, R.J., Scherer, K.R. & Goldsmith, H.H. (Eds.) (2009). *Handbook of Affective Sciences*, Oxford Univ. Press.
- Sander, D. & Scherer, K.R. (Eds.) (2009). *The Oxford Companion to Emotion and the Affective Sciences*, Oxford Univ. Press.

KEY TERMS & DEFINITIONS

Appearance: The appearance of the face consists of the visual features of the facial image.

Shape: The shape is the spatial configuration of features, often represented by a set of named points (i.e. landmarks). It is also called the structural information.

Basic Emotions: According to Ekman, these are ‘happiness’, ‘sadness’, ‘anger’, ‘fear’, ‘surprise’ and ‘disgust’, which have universal manifestations on the face, readable by people regardless of cultural background.

Facial Action Unit: A facial action unit is an objective description of a movement of the face, and serves to identify the unitary components of facial expressions.

Landmark: A fiducial point on the face, also called an anchor point, is a semantically identifiable location that can be identified and tracked in a given facial image. Typical landmarks are mouth and eye corners, tips of the nose and chin, iris centres, etc.