

Face for Ambient Interface

Maja Pantic

Imperial College, Computing Department, 180 Queens Gate,
London SW7 2AZ, U.K.
m.pantic@imperial.ac.uk

Abstract. The human face is used to identify other people, to regulate the conversation by gazing or nodding, to interpret what has been said by lip reading, and to communicate and understand social signals, including affective states and intentions, on the basis of the shown facial expression. Machine understanding of human facial signals could revolutionize user-adaptive social interfaces, the integral part of ambient intelligence technologies. Nonetheless, development of a face-based ambient interface that detects and interprets human facial signals is rather difficult. This article summarizes our efforts in achieving this goal, enumerates the scientific and engineering issues that arise in meeting this challenge and outlines recommendations for accomplishing this objective.

1 Introduction

Films portraying the future often contain visions of human environments of the future. Fitted out with arrays of intelligent, yet invisible devices, homes, transportation means and working spaces of the future can anticipate every need of their inhabitants (Fig. 1). It is this vision of the future that coined the term “ambient intelligence”. According to the Ambient Intelligence (AmI) paradigm, humans will be surrounded by intelligent interfaces that are supported by computing and networking technology embedded in all kinds of objects in the environment and that are sensitive and responsive to the presence of different individuals in a seamless and unobtrusive way [1,40,79]. Thus, AmI involves the convergence of ubiquitous computing, ubiquitous communication, and social user interfaces [64,71] and it assumes a shift in computing – from desktop computers to a multiplicity of smart computing devices diffused into our environment. In turn, it assumes that computing will move to the background, that it will weave itself into the fabric of everyday living spaces and disappear from the foreground [45,74,84], projecting the human user into it [65], and leaving the stage to intuitive social user interfaces. Nonetheless, as computing devices disappear from the scene, become invisible, weaved into our environment, a new set of issues concerning the interaction between ambient intelligence technology and humans is created [74,88]. How can we design the interaction of humans with devices that are invisible? How can we design implicit interaction for sensor-based interfaces? What about users? What does a home dweller, for example, actually want? What are the relevant parameters that can be used by the systems to support us in our activities? If the context is key, how do we arrive at context-aware systems?

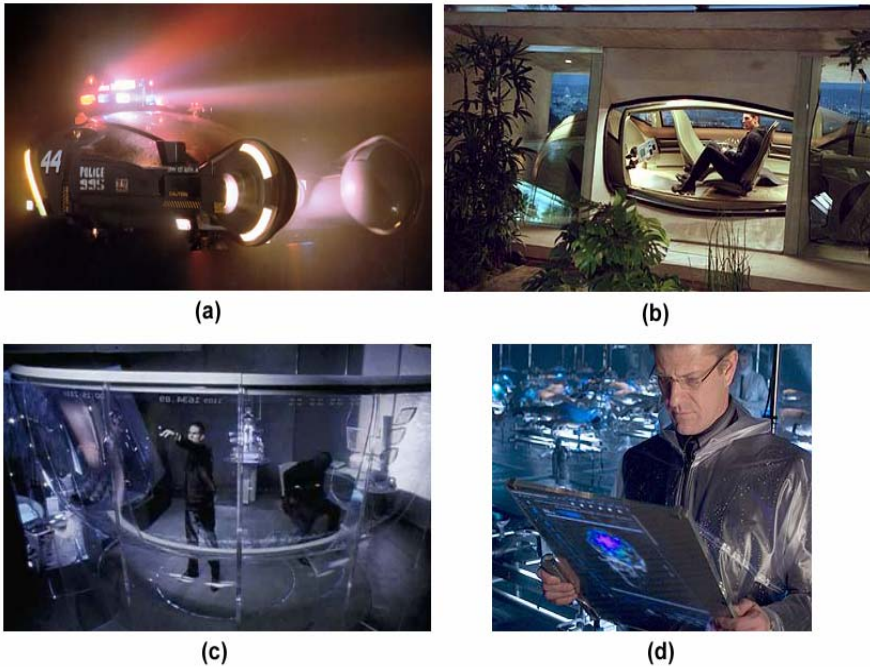


Fig. 1. Human environments of the future envisioned in motion pictures: (a) speech-based interactive car (*Blade Runner*, 1982), (b) speech- and iris-identification driven car (*Minority Report*, 2002), (c) hand-gesture-based interface (*Minority Report*, 2002), (d) multimedia diagnostic chart and an entirely AmI-based environment (*The Island*, 2005)

One way of tackling these problems is to move from computer-centered designs and toward human-centered designs for human computer interaction (HCI). The former usually involve conventional interface devices like keyboards, mice, and visual displays, and assume that the human will be explicit, unambiguous and fully attentive while controlling information and command flow. This kind of interfacing and categorical computing works well for context-independent tasks like making plane reservations and buying and selling stocks. However, it is utterly inappropriate for interacting with each of the (possibly hundreds) computer systems diffused throughout future AmI environments and aimed at improving the quality of life by *anticipating* the users' needs. The key to ambient interfaces is the ease of use - in this case, the ability to unobtrusively sense the user's behavioral cues and to adapt automatically to the particular user behavioral patterns and the context in which the user acts. Thus, instead of focusing on the computer portion of the HCI context, designs for ambient interfaces should focus on the human portion of the HCI context. They should go beyond the traditional keyboard and mouse to include natural, human-like interactive functions including understanding and emulating social signaling. The design of these functions will require explorations of *what* is communicated (linguistic message, non-linguistic conversational signal, emotion, person identification), *how* the information is communicated (the person's facial expression, head movement, tone of voice, hand

and body gesture), *why*, that is, in which context the information is passed on (where the user is, what his current task is, how he/she feels), and *which* (re)action should be taken to satisfy user needs and requirements.

As a first step towards the design and development of such multimodal context-sensitive ambient interfaces, we investigated facial expressions as a potential modality for achieving a more natural, intuitive, and efficient human interaction with computing technology.

1.1 The Human Face

The human face is the site for major sensory inputs and major communicative outputs. It houses the majority of our sensory apparatus: eyes, ears, mouth and nose, allowing the bearer to see, hear, taste and smell. It houses the speech production apparatus and it is used to identify other members of the species, to regulate conversation by gazing or nodding, and to interpret what has been said by lip reading. Moreover, the human face is an accessible “window” into the mechanisms that govern an individual’s emotional and social life. It is our direct and naturally preeminent means of communicating and understanding somebody’s affective state and intentions on the basis of the shown facial expression [38]. Personality, attractiveness, age and gender can also be seen from someone’s face. Thus, the human face is a multi-signal input-output communicative system capable of tremendous flexibility and specificity [24]. In general, it conveys information via four kinds of signals.

1. *Static facial signals* represent relatively permanent features of the face, such as the bony structure, the soft tissue, and the overall proportions of the face. These signals contribute to an individual’s appearance and are usually exploited for person identification.
2. *Slow facial signals* represent changes in the appearance of the face that occur gradually over time, such as the development of permanent wrinkles and changes in skin texture. These signals can be used for assessing the age of an individual. Note that these signals might diminish the distinctness of the facial features and impede recognition of the rapid facial signals.
3. *Artificial signals* are exogenous features of the face, such as glasses and cosmetics. These signals provide additional information that can be used for gender recognition. Note that these signals might obscure facial features or, conversely, might enhance them.
4. *Rapid facial signals* represent temporal changes in neuromuscular activity that may lead to visually detectable changes in facial appearance, including blushing and tears. These (atomic facial) signals underlie *facial expressions*.

All four classes of signals contribute to facial recognition, i.e., person identification. They all contribute to gender recognition, attractiveness assessment, and personality prediction as well. In Aristotle’s time, a theory has been proposed about mutual dependency between static facial signals (physiognomy) and personality: “soft hair reveal a coward, strong chin a stubborn person, and a smile a happy person”¹. Today,

¹ Although this theory is often attributed to Aristotle [4], this is almost certainly not his work (see [4], p. 83).



Fig. 2. Type of messages communicated by rapid facial signals. First row: affective states (anger, surprise, disbelief and sadness). Second row: emblems (wink and thumbs up), illustrators and regulators (head tilt, jaw drop, look exchange, smile), manipulators (yawn).

few psychologists share the belief about the meaning of soft hair and strong chin, but many believe that rapid facial signals (facial expressions) communicate emotions [2,24,38] and personality traits [2]. In fact, among the type of messages communicated by rapid facial signals are the following [23,67]:

1. *affective states and moods*, e.g., joy, fear, disbelief, interest, dislike, frustration,
2. *emblems*, i.e., culture-specific communicators like wink,
3. *manipulators*, i.e., self-manipulative actions like lip biting and yawns,
4. *illustrators*, i.e., actions accompanying speech such as eyebrow flashes,
5. *regulators*, i.e., conversational mediators such as the exchange of a look, head nods and smiles.

Given the significant role of the face in our emotional and social lives, it is not surprising that the potential benefits of efforts to automate the analysis of facial signals, in particular rapid facial signals, are varied and numerous, especially when it comes to computer science and technologies brought to bear on these issues. As far as natural interfaces between humans and computers (PCs / robots / machines) are concerned, facial expressions provide a way to communicate basic information about needs and demands to the machine. In fact, automatic analysis of rapid facial signals seems to have a natural place in various vision sub-systems, including automated tools for gaze and focus of attention tracking, lip reading, bimodal speech processing, face / visual speech synthesis, and face-based command issuing. Where the user is looking (i.e., gaze tracking) can be effectively used to free computer users from the classic keyboard and mouse. Also, certain facial signals (e.g., a wink) can be associated with

certain commands (e.g., a mouse click) offering an alternative to traditional keyboard and mouse commands. The human capability to “hear” in noisy environments by means of lip reading is the basis for bimodal (audiovisual) speech processing that can lead to the realization of robust speech-driven interfaces. To make a believable “talking head” (avatar) representing a real person, tracking the person’s facial signals and making the avatar mimic those using synthesized speech and facial expressions is compulsory. Combining facial expression spotting with facial expression interpretation in terms of labels like “did not understand”, “disagree”, “inattentive”, and “approves” could be employed as a tool for monitoring human reactions during videoconferences and web-based lectures. Attendees’ facial expressions will inform the speaker (teacher) of the need to adjust the (instructional) presentation.

The focus of the relatively recently-initiated research area of *affective computing* lies on sensing, detecting and interpreting human affective states and devising appropriate means for handling this affective information in order to enhance current HCI designs [61]. The tacit assumption is that in many situations human-machine interaction could be improved by the introduction of machines that can adapt to their users (think about computer-based advisors, virtual information desks, on-board computers and navigation systems, pacemakers, etc.). The information about when the existing processing should be adapted, the importance of such an adaptation, and how the processing/reasoning should be adapted, involves information about how the user feels (e.g. confused, irritated, frustrated, interested). As facial expressions are our direct, naturally preeminent means of communicating emotions, machine analysis of facial expressions forms an indispensable part of affective HCI designs [52].

Automatic assessment of boredom, fatigue, and stress, will be highly valuable in situations where firm attention to a crucial, but perhaps tedious task is essential, such as aircraft and air traffic control, space flight and nuclear plant surveillance, or simply driving a ground vehicle like a truck, train, or car. If these negative affective states could be detected in a timely and unobtrusive manner, appropriate alerts could be provided, preventing many accidents from happening. Automated detectors of fatigue, depression and anxiety could form another step toward personal wellness technologies [20], which scale with the needs of an aging population, as the current medical practices that rely heavily on expensive and overburdened doctors, nurses, and physicians will not be possible any longer. An advantage of machine monitoring is that human observers need not be present to perform privacy-invading monitoring; the automated tool could provide advice, feedback and prompts for better performance based on the sensed user’s facial expressive behavior.

Monitoring and interpretation of facial signals are also important to lawyers, police, security and intelligence agents, who are often interested in issues concerning deception and attitude. Automated facial reaction monitoring could form a valuable tool in these situations, as now only informal interpretations are used. Systems that can recognize friendly faces or, more important, recognize unfriendly or aggressive faces, determine an unwanted intrusion or hooligan behavior, and inform the appropriate authorities, represent another application of facial measurement technology. Systems that adjust music and light levels according to the number, activity, and mood of the users form also an AmI application of this technology.

1.2 Why Face for Ambient Interface?

One can easily formulate the answer to this question by considering the breadth of the applied research on AmI and perceptual HCI that uses measures of the face and facial behavior. The preceding section has separately enumerated several computer science research areas and multiple applications in healthcare, industrial, commercial, and professional sectors that would reap substantial benefits from facial measurement technology. This section emphasizes these benefits in the light of the design guidelines defined for ambient interfaces (Table 1).

Table 1. The fitness of facial measurement technology for the design of ambient interfaces based upon the design guidelines defined for such interfaces [30,63]

<i>Effective</i>	One basic goal for ambient interfaces is continuous provision of background information without disrupting user's foreground tasks. Monitoring the user's attentiveness to the current foreground task based upon his/her facial behavior could help realizing this goal.
<i>Efficient</i>	Ambient interfaces should support users in carrying out their tasks efficiently. Examples of how facial measurement technology can help achieving this goal include face-based user identification that relieves users from typing user names and passwords, adapting the amount of the presented information to the level of user's fatigue, and provision of appropriate assistance if confusion can be read from the user's face.
<i>easy to learn & remember</i>	It is particularly challenging to achieve ambient interfaces that are easy for the users to learn and to remember, since novel metaphors are used. Nevertheless, the face is the human natural means used to regulate the interaction by gazing, nodding, smiling, frowning, etc. Face-based interfaces would probably be the easiest for the users to "learn and remember".
<i>context-aware</i>	One basic goal for ambient interfaces is the achievement of the systems' awareness of the context in which the users act [59]: who they are, what their current task is, where they are, how they feel. Face recognition, gaze tracking, and facial affect analysis offer the basis for the design of personalized, affective, task-dependent, natural feeling interfaces.
<i>Control adequate</i>	It is a particular challenge to realize unambiguous mapping between controls and their effects in the case of ambient interfaces. Note, however, that there should be little (if any) ambiguity about the effect of input facial signals like identity, gaze focus, smiles and frowns, especially when it comes to typically constrained AmI scenarios involving a certain individual's home, car, or work space.
<i>Domain adequate</i>	Adequacy of the ambient interfaces for the target domain including the users, their tasks and the environment should be ensured [73]. Although facial measurement technology cannot ensure realization of this goal on its own, it has the potential to accommodate a broad range of users through customized face-based interaction controls for support of different users with different abilities and needs.

participatory design In the design of ambient interfaces, it is important to stimulate the users to contribute to the design at early stages, so that products tailored to their preferences, needs and habits can be ensured. This is inherent in face-based AmI technology, which relies on machine learning and sees the human user as the main actor. Improving the quality of life by anticipating one's needs via rigid, impersonalized systems is unrealistic; the necessary mapping of sensed facial signals (there are more than 7000 of these [69]) onto a set of controls and preemptive behaviors is by far too complex to be precompiled and hardwired into the system. Systems should learn their expertise by having the user instruct them (explicitly / implicitly) on the desired (context-sensitive) interpretations of sensed facial signals [52].

In addition, ambient interfaces should have *good utility* (e.g. they are not suitable for complex information processing) and be *transparent* (i.e. users should be aware, at any time, of what is expected from them, whether the input was received, whether the actions are or will be performed, etc.). Facial measurement technology represents a novel interface modality; it has no direct answers to these basic design questions.

While all agree that facial measurement technology has a natural place in AmI technologies, especially in human-centered natural-feeling ambient interfaces, one should be aware of the likelihood that face-based ambient interfaces still lie in the relatively distant future. Although humans detect and analyze faces and facial expressions in a scene with little or no effort, development of an automated system that accomplishes this task is rather difficult. There are several related problems [52]. The first is to find faces in the scene independent of clutter, occlusions, and variations in head pose and lighting conditions. Then, facial features such as facial characteristic points (e.g. the mouth corners) or parameters of a holistic facial model (e.g. parameters of a fitted Active Appearance Model) should be extracted from the regions of the scene that contain faces. The system should perform this accurately, in a fully automatic manner and preferably in real time. Eventually, the extracted facial information should be interpreted in terms of facial signals (identity, gaze direction, winks, blinks, smiles, affective states, moods) in a context-dependent (personalized, task- and application-dependent) manner. For exhaustive surveys of the entire problem domain, the readers are referred to: Samal and Iyengar [68] for an overview of early works, Tian et al. [78] and Pantic [47] for surveys of techniques for detecting facial muscle actions (AUs), and Pantic and Rothkrantz [51,52] for surveys of current efforts. These surveys indicate that although the fields of computer vision and facial information processing witnessed rather significant advances in the past few years, most of the aforementioned problems still represent significant challenges facing the researchers in these and the related fields. This paper summarizes our efforts in solving some of these problems, enumerates the scientific and engineering issues that arise in meeting these challenges and outlines recommendations for accomplishing the new facial measurement technology.

2 Face Detection

The first step in facial information processing is face detection, i.e., identification of all regions in the scene that contain a human face. The problem of *finding faces* should be solved regardless of clutter, occlusions, and variations in head pose and lighting conditions. The presence of non-rigid movements due to facial expression and a high degree of variability in facial size, color and texture make this problem even more difficult. Numerous techniques have been developed for face detection in still images [39,87]. However, most of them can detect only upright faces in frontal or near-frontal view. The efforts that had the greatest impact on the community (as measured by, e.g., citations) include the following.

Rowley et al. [66] used a multi-layer neural network to learn the face and non-face patterns from the intensities and spatial relationships of pixels in face and non-face images. Sung and Poggio [75] have proposed a similar method. They used a neural network to find a discriminant function to classify face and non-face patterns using distance measures. Moghaddam and Pentland [44] developed a probabilistic visual learning method based on density estimation in a high-dimensional space using an eigenspace decomposition. The method has been applied to face localization, coding and recognition. Pentland et al. [60] developed a real-time, view-based and modular (by means of incorporating salient features, such as the eyes and the mouth) eigenspace description technique for face recognition in variable pose.

Among all the face detection methods that have been employed by automatic facial expression analyzers, the most significant work is arguably that of Viola and Jones [82]. They developed a real-time face detector consisting of a cascade of classifiers trained by AdaBoost. Each classifier employs integral image filters, which remind of Haar Basis functions and can be computed very fast at any location and scale (Fig. 4(a)). This is essential to the speed of the detector. For each stage in the cascade, a subset of features is chosen using a feature selection procedure based on AdaBoost.

There are several adapted versions of the Viola-Jones face detector and the one that we employ in our systems uses GentleBoost instead of AdaBoost. It also refines the originally proposed feature selection by finding the best performing single-feature classifier from a new set of filters generated by shifting and scaling the chosen filter by two pixels in each direction, as well as by finding composite filters made by reflecting each shifted and scaled feature horizontally about the center and superimposing it on the original [27]. Finally the employed version of the face detector uses a smart training procedure in which, after each single feature, the system can decide whether to test another feature or to make a decision. By this, the system retains information about the continuous outputs of each feature detector rather than converting to binary decisions at each stage of the cascade. The employed face detector was trained on 5000 faces and millions of non-face patches from about 8000 images collected from the web by Compaq Research Laboratories [27]. On the test set of 422 images from the Cohn-Kanade facial expression database [37], the most commonly used database of face images in the research on facial expression analysis, the detection rate was 100% [83].

3 Facial Feature Extraction

After the presence of a face has been detected in the observed scene, the next step is to extract the information about the displayed facial signals. The problem of *facial feature extraction* from regions in the scene that contain a human face may be divided into at least three dimensions [51]:

1. Is temporal information used?
2. Are the features holistic (spanning the whole face) or analytic (spanning sub-parts of the face)?
3. Are the features view- or volume based (2D/3D)?

Given this glossary and if the goal is face recognition, i.e., identifying people by looking at their faces, most of the proposed approaches adopt 2D holistic static facial features. On the other hand, many approaches to automatic facial expression analysis adopt 2D analytic spatio-temporal facial features [52]. This finding is also consistent with findings from the psychological research suggesting that the brain processes faces holistically rather than locally whilst it processes facial expressions locally [9,13]. What is, however, not entirely clear yet is whether information on facial expression is passed to the identification process to aid recognition of individuals or not. Some experimental data suggest this [42]. Although relevant for the discussion of facial measurement tools and face-based ambient interfaces, these issues are not elaborated further in this paper, as the focus of our past research was mainly automatic facial expression analysis. For exhaustive surveys of efforts aimed at face recognition, the readers are referred to: Zhao et al. [90], Bowyer [12], and Li and Jain [39].

Most of the existing facial expression analyzers are directed toward 2D spatio-temporal facial feature extraction, including the methods proposed by our research team. The usually extracted facial features are either *geometric features*, such as the shapes of the facial components (eyes, mouth, etc.) and the locations of facial fiducial points (corners of the eyes, mouth, etc.) or *appearance features* representing the texture of the facial skin, including wrinkles, bulges, and furrows [7,8]. Typical examples of geometric-feature-based methods are those of Gokturk et al. [29], who used 19 point face mesh, and of Pantic et al. [49,53,81], who used a set of facial characteristic points like the ones illustrated in Fig. 10. Typical examples of *hybrid*, geometric- and appearance-feature-based methods are those of Tian et al. [77], who used shape-based models of eyes, eyebrows and mouth and transient features like crows-feet wrinkles and nasolabial furrow, and of Zhang and Ji [89], who used 26 facial points around the eyes, eyebrows, and mouth and the same transient features as Tian et al [77]. Typical examples of appearance-feature-based methods are those of Bartlett et al. [8,21] and Guo and Dyer [32], who used Gabor wavelets, of Anderson and McOwen [3], who used a holistic, monochrome, spatial-ratio face template, and of Valstar et al. [80], who used temporal templates (see Section 3.2).

It has been reported that methods based on geometric features are usually outperformed by those based on appearance features using, e.g., Gabor wavelets or eigen-faces [7]. Recent studies have shown that this claim does not always hold [49,81]. Moreover, it seems that using both geometric and appearance features might be the best choice in the case of certain facial expressions [49].

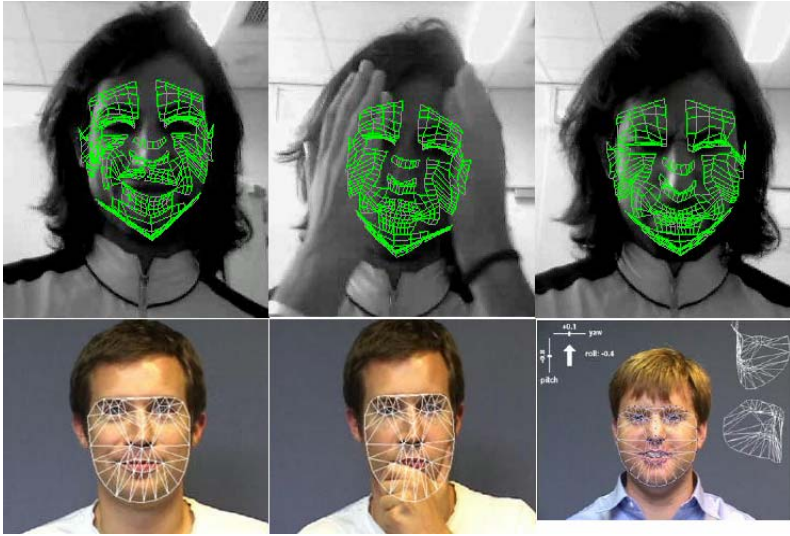


Fig. 3. Examples of 2D and 3D face models. First row: Results of Tao-Huang 3D-wireframe face-model fitting algorithm for happy, occluded and angry face image frames [76]. Second row: Results of the CMU 2D-AAM fitting algorithm for happy and occluded face image frames and results of fitting the CMU 2D+3D AAM [5,85].

Few approaches to automatic facial expression analysis based on 3D face modeling have been proposed recently (Fig. 3). Gokturk et al. [29] proposed a method for recognition of facial signals like brow flashes and smiles based upon 3D deformations of the face tracked on stereo image streams using a 19-point face mesh and standard optical flow techniques. The work of Cohen et al. [14] focuses on the design of Bayesian network classifiers for emotion recognition from face video based on facial features tracked by so-called Piecewise Bezier Volume Deformation tracker [76]. This tracker employs an explicit 3D wireframe model consisting of 16 surface patches embedded in Bezier volumes. Cohn et al. [15] focus on automatic analysis of brow actions and head movements from face video and use a cylindrical head model to estimate the 6 degrees of freedom of head motion. Baker and his colleagues developed several algorithms for fitting 2D and combined 2D+3D Active Appearance Models to images of faces [85], which can be used further for various studies concerning human facial behavior [5]. 3D face modeling is highly relevant to the present goals due to its potential to produce view-independent facial signal recognition systems. The main shortcomings of the current methods concern the need of a large amount of manually annotated training data and an almost always required manual selection of landmark facial points in the first frame of the video-based input on which the face model will be warped to fit the face. Automatic facial feature point detection of the kind proposed in Section 3.1 offers a solution to these problems.

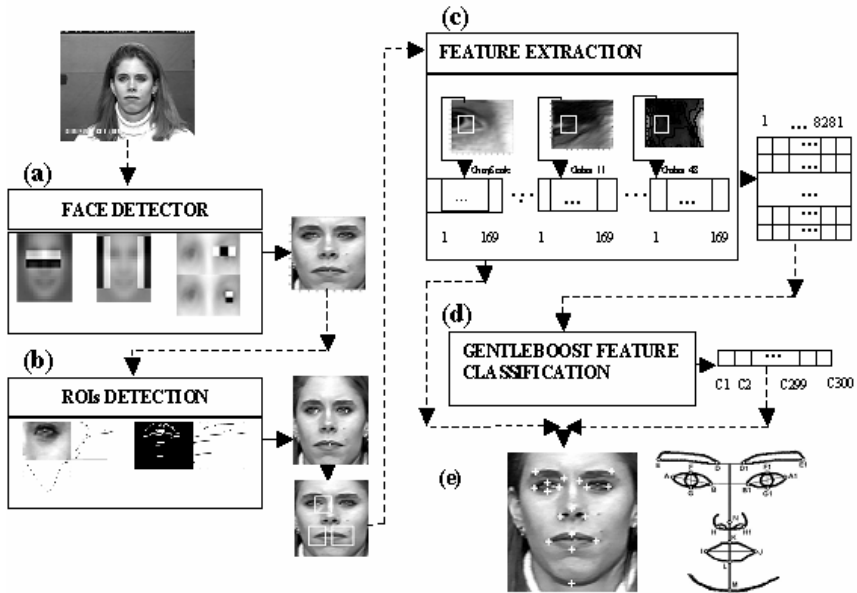


Fig. 4. Outline of our Facial Point Detection method [83]. (a) Face detection using Haar-feature-based GentleBoost classifier [27]; (b) ROI extraction, (c) feature extraction based on Gabor filtering, (d) feature selection and classification using GentleBoost classifier, (e) output of the system compared to the face drawing with facial landmark points we aim to detect.

3.1 Geometric Feature Extraction – Facial Feature Point Detection

The method that we use for fully automatic detection of 20 facial feature points, illustrated in Fig. 4(e) and Fig 10, uses Gabor-feature-based boosted classifiers [83]. The method adopts the fast and robust face detection algorithm explained in Section 2, which represents an adapted version of the original Viola-Jones detector [27,82].

The detected face region is then divided in 20 regions of interest (ROIs), each one corresponding to one facial point to be detected. The irises and the medial point of the mouth are detected first. The detection is done through a combination of heuristic techniques based on the analysis of the vertical and horizontal histograms of the upper and the lower half of the face-region image achieves this (Fig. 5). Subsequently, we use the detected positions of the irises and the medial point of the mouth to localize 20 ROIs. An example of ROIs extracted from the face region for points B, I, and J, is depicted in Fig. 4(b).

The employed facial feature point detection method uses individual feature patch templates to detect points in the relevant ROI. These feature models are GentleBoost templates built from both gray level intensities and Gabor wavelet features. Recent work has shown that a Gabor approach for local feature extraction outperformed Principal Component Analysis (PCA), the Fisher’s Linear Discriminant (FLD) and the Local Feature Analysis [21]. This finding is also consistent with our experimental data that show the vast majority of features (over 98%) that were selected by the

utilized GentleBoost classifier [28] were from the Gabor filter components rather than from the gray level intensities. The essence of the success of Gabor filters is that they remove most of the variability in image due to variation in lighting and contrast, while being robust against small shifts and deformation [35].

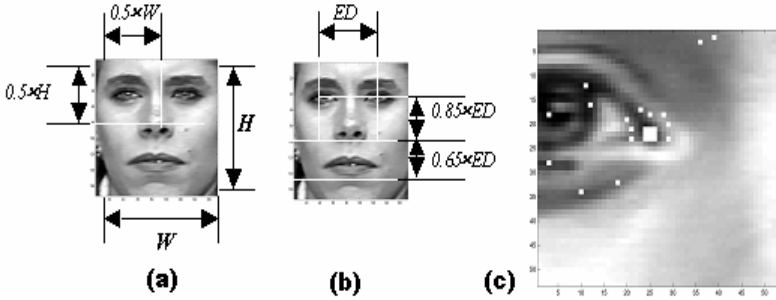


Fig. 5. (a) Dividing the face horizontally in half and dividing the upper face region vertically in half. (b) Finding the mouth region within the face region by means of Eye Distance (ED). (c) Positive and negative examples for training point B. The big white square on the inner corner of the eye represents 9 positive examples. Around that square are 8 negative examples randomly chosen near the positive examples. Another 8 negative examples are randomly chosen from the rest of the region.

The feature vector for each facial point is extracted from the 13×13 pixels image patch centered on that point. This feature vector is used to learn the pertinent point's patch template and, in the testing stage, to predict whether the current point represents a certain facial point or not. This 13×13 pixels image patch is extracted from the gray scale image of the ROI and from 48 representations of the ROI obtained by filtering the ROI with a bank of 48 Gabor filters at 8 orientations and 6 spatial frequencies (2:12 pixels/cycle at $\frac{1}{2}$ octave steps). Thus, $169 \times 49 = 8281$ features are used to represent one point. Each feature contains the following information: (i) the position of the pixel inside the 13×13 pixels image patch, (ii) whether the pixel originates from a grayscale or from a Gabor filtered representation of the ROI, and (iii) if appropriate, which Gabor filter has been used (Fig. 4(c)).

In the training phase, GentleBoost feature templates are learned using a representative set of positive and negative examples. As positive examples for a facial point, we used 9 image patches centered on the true point and on 8 positions surrounding the true (manually labeled) facial point in a training image. For each facial point we used two sets of negative examples. The first set contains 8 image patches randomly displaced 2-pixels distance from any of the positive examples. The second set contains 8 image patches randomly displaced in the relevant ROI (Fig. 5).

In the testing phase, each ROI is filtered first by the same set of Gabor filters used in the training phase (in total, 48 Gabor filters are used). Then, for a certain facial point, an input 13×13 pixels window (*sliding window*) is slid pixel by pixel across 49 representations of the relevant ROI (grayscale plus 48 Gabor filter representations). For each position of the sliding window, the GentleBoost classification method [28]

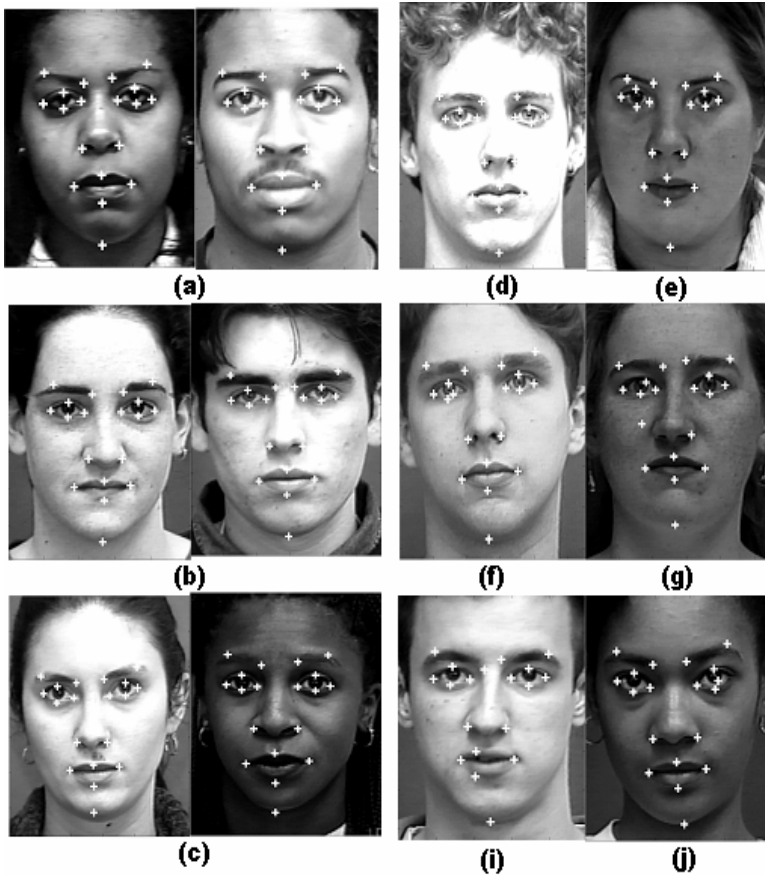


Fig. 6. Typical results of our Facial Feature Point Detector [83]. (a)-(c) Examples of accurate detection. (d)-(j) Examples of inaccurate detection of point D1 (d), point E (e), point B (f), points F and H (g), point D and K (i), and point A1 (j). For point notation see Fig. 12.

outputs a response depicting the similarity between the 49-dimensional representation of the sliding window and the learned feature point model. After scanning the entire ROI, the position with the highest response reveals the feature point in question.

We trained and tested the facial feature detection method on the Cohn-Kanade facial expression database [37]. We used only the first frames of the 300 Cohn-Kanade database samples. No further registration of the images was performed. The 300 images of the data set were divided into 3 subsets containing 100 images each. The proposed method has been trained and tested using a leave-one-subset-out cross validation. To evaluate the performance of the method, each of the automatically located facial points was compared to the true (manually annotated) point. As explained above, we used as positive examples the true location of the point and 8 positions surrounding the true facial point in a training image. Hence, automatically detected

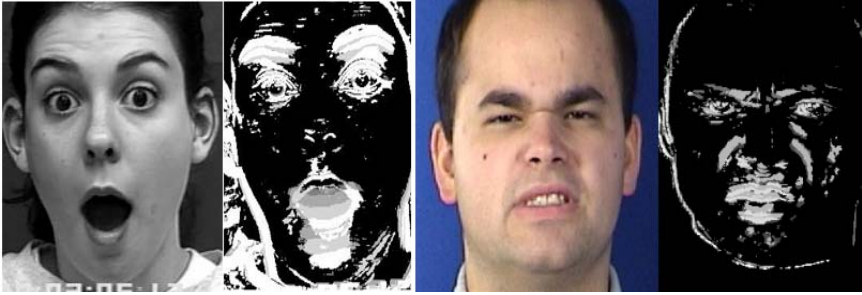


Fig. 7. Original maximal activation (apex) frames and the related Motion History Images (MHI). Left: AU1+AU2+AU5+AU27 (from the Cohn-Kanade facial expression database). Right: AU9+AU10+AU25 (from the MMI facial expression database).

points displaced 1-pixel distance from relevant true facial points are regarded as SUCCESS. Additionally, we define errors with respect to the inter-ocular distance measured in the test image (80 to 120 pixels in the case of image samples from the Cohn-Kanade database). An automatically detected point displaced in any direction, horizontal or vertical, less than 5% of inter-ocular distance (i.e., 4 to 6 pixels in the case of image samples from the Cohn-Kanade database) from the true facial point is regarded as SUCCESS. This is in contrast to the current related approaches developed elsewhere (e.g. [17]), which are usually regarded as SUCCESS if the bias of automatic labeling result to the manual labeling result is less than 30% of the true (annotated manually) inter-ocular distance.

Overall, we achieved an average recognition rate of 93% for 20 facial feature points using the above described evaluation scheme. Typical results are shown in Fig. 6. Virtually all misclassifications (most often encountered with points F1 and M) can be attributed to the lack of consistent rules for manual annotation of the points. For details about this method, the readers are referred to Vukadinovic and Pantic [83].

3.2 Appearance Feature Extraction – Temporal Templates

Temporal templates are 2D images constructed from image sequences, which show motion history, that is, where and when motion in the input image sequence has occurred [11]. More specifically, the value of a pixel in a Motion History Image (MHI) decays over time, so that a high intensity pixel denotes recent motion, a low intensity pixel denotes a motion that occurred earlier in time, and intensity zero denotes no motion at all at that specific location (Fig. 7). A drawback innate to temporal templates proposed originally by Bobick and Davis [11] is the problem of motion self-occlusion due to overwriting. Let us explain this problem by giving an example. Let us denote an upward movement of the eyebrows as action A_1 and a downward movement of the eyebrows back to the neutral position as action A_2 . Both actions produce apparent motion in the facial region above the neutral position of the eyebrows (Fig. 7). If A_2 follows A_1 in time and if the motion history of both actions is recorded within a single Motion History Image (MHI), then the motion history of action A_2 overwrites the motion history of A_1 ; the information about the motion

history of action A_i is lost. To overcome this problem, we proposed to record the motion history at multiple time intervals and to construct Multilevel Motion History Image (MMHI), instead of recording the motion history once for the entire image sequence and constructing a single MHI [80].

Before we can construct a MMHI from an input video, the face present in the video needs to be registered in two ways. Intra registration removes all rigid head movements within the input video, while the inter registration places the face at a predefined location in the scene. This transformation uses facial points whose spatial position remains the same even if a facial muscle contraction occurs (i.e., points B, B1, and N illustrated in Fig. 10). The inter registration process warps the face onto a predefined “normal” face, eliminating inter-person variation of face shape and facilitating the comparison between the facial expression shown in the input video and template facial expressions. Under the assumption that each input image sequence begins and ends with a neutral facial expression, we downsample the number of frames to a fixed number of $(n+1)$ frames. In this way, our system becomes robust to the problem of varying duration of facial expressions.

After the registration and time warping of the input image sequence, the MHI is obtained as follows. Let $I(x, y, t)$ be an image sequence of pixel intensities of k frames and let $D(x, y, t)$ be the binary image that results from pixel intensity change detection, that is by thresholding $|I(x, y, t) - I(x, y, t-1)| > th$, where x and y are the spatial coordinates of picture elements and th is the minimal intensity difference between two images. In an MHI, say H_t , the pixel intensity is a function of the temporal history of motion at that point with t being a frame of the downsampled input video (with $(n+1)$ frames). Using the known parameter n , H_t is defined as:

$$H_t(x, y, t) = \begin{cases} s * t & D(x, y, t) = 1 \\ H_t(x, y, t-1) & \text{otherwise} \end{cases} \quad (1)$$

where $s = (255/n)$ is the intensity step between two history levels and where $H_t(x, y, t) = 0$ for $t \leq 0$. The final MHI, say $H(x, y)$, is found by iteratively computing equation (1) for $t = 1 \dots n+1$.

With an MMHI, we want to encode motion occurring at different time instances on the same location such that it is uniquely decodable later on. To do so, we use a simple bit-wise coding scheme. If motion occurs at time instance t at position (x, y) , we add 2 to the power of $(t-1)$ to the old value of the MMHI:

$$M(x, y, t) = M(x, y, t-1) + D(x, y, t) \cdot 2^{t-1} \quad (2)$$

with $M(x, y, t) = 0$ for $t \leq 0$. Because of the bitwise coding scheme, we are able to separate multiple motions occurring at the same position in the classification stage.

We utilized further a temporal-template-based face image sequence representation for automatic recognition of facial signals such as brow flashes, smiles, frowns, etc. (i.e., for AU detection). Comparison of two classification schemes: (i) a two-stage classifier combining a kNN-based and a rule-based classifier, and (ii) a SNoW classifier, can be found in Valstar et al. [80]. The evaluations studies on two different databases, the Cohn-Kanade [37] and the MMI facial expression database [56], suggest that (M)MHIs are very suitable for detecting various facial signals. Especially

AU1+AU2 (eyebrows raised), AU10+AU25 (raised upper lip), AU12+AU25 (smile with lips parted) and AU27 (mouth stretched vertically) are easily recognized. However, as it is the case with all template-based methods, for each and every facial signal (new class) to be recognized, a separate template should be learned. Given that there are more than 7000 different facial expressions [69], template-based methods including temporal templates, do not represent the best choice for realizing facial measurement tools.

4 Facial Feature Tracking

Contractions of facial muscles induce movements of the facial skin and changes in the appearance of facial components such as eyebrows, nose, and mouth. Since motion of the facial skin produces optical flow in the image, a large number of researchers have studied optical flow tracking [39,51]. The optical flow approach to describing face motion has the advantage of not requiring a facial feature extraction stage of processing. Dense flow information is available throughout the entire facial area, regardless of the existence of facial components, even in the areas of smooth texture such as the cheeks and the forehead. Because optical flow is the visible result of movement and is expressed in terms of velocity, it can be used to represent facial actions directly. One of the first efforts to utilize optical flow for recognition of facial expressions was the work of Mase [43]. Many other researchers adopted this approach including Black and Yacoob [10], who used the flows within local facial areas of the facial components for expression recognition purposes. For exhaustive surveys of these methods, the reader is referred to Pantic and Rothkrantz [51] and Li and Jain [39].

Standard optical flow techniques [6,41,72] are also most commonly used for tracking facial feature points. DeCarlo and Metaxas [19] presented a model-based tracking algorithm in which face shape model and motion estimation are integrated using optical flow and edge information. Gokturk et al. [29] track the points of their 19-point face mesh on the stereo image streams using the standard Lucas-Kanade optical flow algorithm [41]. To achieve facial feature point tracking Tian et al. [77] and Cohn et al. [15,16] used the standard Lucas-Kanade optical flow algorithm too. To realize fitting of 2D and combined 2D+3D Active Appearance Models to images of faces [85], Xiao et al. use an algorithm based on an "inverse compositional" extension to the Lucas-Kanade algorithm.

To omit the limitations inherent in optical flow techniques, such as the accumulation of error and the sensitivity to noise, occlusion, clutter, and changes in illumination, several researchers used sequential state estimation techniques to track facial feature points in image sequences. Both, Zhang and Ji [89] and Gu and Ji [31] used facial point tracking based on a Kalman filtering scheme, which is the traditional tool for solving sequential state problems. The derivation of the Kalman filter is based on a state-space model [36], governed by two assumptions: (i) linearity of the model and (ii) Gaussianity of both the dynamic noise in the process equation and the measurement noise in the measurement equation. Under these assumptions, derivation of the Kalman filter leads to an algorithm that propagates the mean vector and covariance matrix of the state estimation error in an iterative manner and is optimal in the Bayesian setting. To deal with the state estimation in nonlinear dynamical systems, the extended Kalman filter has been proposed, which is derived through linearization of the

state-space model. However, many of the state estimation problems, including human facial expression analysis, are nonlinear and quite often non-Gaussian too. Thus, if the face undergoes a sudden or rapid movement, the prediction of features positions from Kalman filtering will be significantly off. To handle this problem, Zhang and Ji [89] and Gu and Ji [31] used the information about the IR-camera-detected pupil location together with the output of Kalman filtering to predict facial features positions in the next frame of an input face video. To overcome these limitations of the classical Kalman filter and its extended form in general, particle filters have been proposed. An extended overview of the various facets of particle filters can be found in [33].

The tracking scheme that we utilize to track facial feature points in an input face image sequence is based on particle filtering. The main idea behind particle filtering is to maintain a set of solutions that are an efficient representation of the conditional probability $p(\alpha|Y)$, where α is the state of a temporal event to be tracked given a set of noisy observations $Y = \{y^1, \dots, y^-, y\}$ up to the current time instant. This means that the distribution $p(\alpha|Y)$ is represented by a set of pairs $\{(s_k, \pi_k)\}$ such that if s_k is chosen with probability equal to π_k , then it is as if s_k was drawn from $p(\alpha|Y)$. By maintaining a set of solutions instead of a single estimate (as is done by Kalman filtering), particle filtering is able to track multimodal conditional probabilities $p(\alpha|Y)$, and it is therefore robust to missing and inaccurate data and particularly attractive for estimation and prediction in nonlinear, non-Gaussian systems. In the particle filtering framework, our knowledge about the *a posteriori* probability $p(\alpha|Y)$ is updated in a recursive way. Suppose that at a previous time instance we have a particle-based representation of the density $p(\alpha^-|Y^-)$, i.e., we have a collection of K particles and their corresponding weights (i.e. $\{(s_k^-, \pi_k^-)\}$). Then, the classical particle filtering algorithm, so-called Condensation algorithm, can be summarized as follows [34].

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

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

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

To track facial feature points for the purposes of facial expression analysis, we utilize two different extensions to classical Condensation algorithm. The first one is the Auxiliary Particle Filtering introduced by Pitt and Shepard [62], which addresses the first drawback of the Condensation algorithm by favoring particles that end up in areas with high likelihood when propagated with the transition density $p(\alpha | \alpha^-)$. The second extension to classical Condensation algorithm that we utilize for facial point tracking is the Particle Filtering with Factorized Likelihoods proposed by Patras and Pantic [57]. This algorithm addresses all of the aforementioned problems inherent in the Condensation algorithm by extending the Auxiliary Particle Filtering to take into account the interdependencies between the different parts of the state α . In order to do so we partition the state α into sub-states α_i that correspond to the different facial features, that is $\alpha = \langle \alpha_1, \dots, \alpha_n \rangle$. At each frame of the sequence we obtain a particle-based representation of $p(\alpha | y)$ in two stages. In the first stage, we apply one complete step of a particle filtering algorithm (in our case the auxiliary particle filtering) in order to obtain a particle-based representation of $p(\alpha_i | y)$, for each facial feature i . That is, at the first stage, each facial feature is tracked for one frame independently from the other facial features. At the second stage, interdependencies between the sub-states are taken into consideration, in a scheme that samples complete particles from the proposal distribution $\prod_i p(\alpha_i | y)$ and evaluates them using $p(\alpha | \alpha^-)$. The density $p(\alpha | \alpha^-)$, that captures the interdependencies between the locations of the



Fig. 8. Results of the facial point tracking in face-profile image sequences [49]. First row: frames 1 (neutral), 48 (onset AU29), 59 (apex AU29), and 72 (offset AU29). Second row: frames 1 (neutral), 25 (onset AU12), 30 (onset AU6+12), and 55 (apex AU6+12+25+45).

facial features is estimated using a kernel-based density estimation scheme. Finally, we define an observation model that is based on a robust color-based distance between the color template $o = \{o_i | i = 1 \dots M\}$ and a color template $c = \{c_i | i = 1 \dots M\}$ at the current frame. We attempt to deal with shadows by compensating for the global intensity changes. We use the distance function d , defined by equation (3), where M is the number of pixels in each template, m_c is the average intensity of template $c = \{c_i\}$, m_o is the average intensity of template $o = \{o_i\}$, i is the pixel index, and $\rho(\cdot)$ is a robust error function such as the Geman-McClure.

$$d = \sum_{i=1}^M \rho \left(\left\| \frac{c_i}{m_c} - \frac{o_i}{m_o} \right\|_1 \mu_c \right) / M . \quad (3)$$

Typical results of the Auxiliary Particle Filtering, adapted for the problem of color-based template tracking as explained above and applied for tracking facial points in video sequences of profile view of the face [49], are shown in Fig. 8. Typical results of the Particle Filtering with Factorized Likelihoods, applied for tracking color-based templates of facial points in image sequences of faces in frontal-view [81], are shown in Fig. 9.



Fig. 9. Results of the facial point tracking in frontal-view face image sequences [81]. First row: frames 1 (neutral), 14 (onset AU1+2+5+20+25+26), and 29 (apex AU1+2+5+20+25+26). Second row: frames 1 (neutral), 32 (apex AU4+7+17+24), and 55 (apex AU45, offset AU4+7+17+24).

5 Facial Action Coding

Most approaches to automatic facial expression analysis attempt to recognize a small set of prototypic emotional facial expressions, i.e., fear, sadness, disgust, happiness, anger, and surprise (e.g. [3,10,14,32,42,43,89]; for exhaustive surveys of the past work on this research topic, the reader is referred to [51,52,68]). This practice may follow from the work of Darwin [18] and more recently Ekman [22,24,38], who suggested that basic emotions have corresponding prototypic facial expressions. In everyday life, however, such prototypic expressions occur relatively rarely; emotions are displayed more often by subtle changes in one or few discrete facial features, such as the raising of the eyebrows in surprise [46]. To detect such subtlety of human emotions and, in general, to make the information conveyed by facial expressions available for usage in various applications summarized in Sections 1.1 and 1.2, automatic recognition of rapid facial signals, i.e., facial muscle actions, such as the action units (AUs) of the Facial Action Coding System (FACS) [25,26], is needed.

5.1 Facial Action Coding System























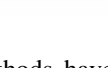
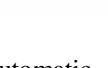
Rapid facial signals are movements of the facial muscles that pull the skin, causing a temporary distortion of the shape of the facial features and of the appearance of folds, furrows, and bulges of skin. The common terminology for describing rapid facial signals refers either to culturally dependent linguistic terms indicating a specific change in the appearance of a particular facial feature (e.g., smile, smirk, frown, sneer) or to the linguistic universals describing the activity of specific facial muscles that caused the observed facial appearance changes.

There are several methods for linguistically universal recognition of facial changes based on the facial muscular activity [69]. From those, the facial action coding system (FACS) proposed by Ekman et al. [25,26] is the best known and most commonly used system. It is a system designed for human observers to describe changes in the facial expression in terms of visually observable activations of facial muscles. The changes in the facial expression are described with FACS in terms of 44 different Action Units (AUs), each of which is anatomically related to the contraction of either a specific facial muscle or a set of facial muscles. Examples of different AUs are given in Table 2. Along with the definition of various AUs, FACS also provides the rules for visual detection of AUs and their temporal segments (onset, apex, offset) in a face image. Using these rules, a FACS coder (that is a human expert having a formal training in using FACS) decomposes a shown facial expression into the AUs that produce the expression.

5.2 Automated Facial Action Coding

Although FACS provides a good foundation for AU-coding of face images by human observers, achieving AU recognition by a computer is by no means a trivial task. A problematic issue is that AUs can occur in more than 7000 different complex combinations [69], causing bulges (e.g., by the tongue pushed under one of the lips) and various in- and out-of-image-plane movements of permanent facial features (e.g., jetted jaw) that are difficult to detect in 2D face images.

Table 2. Examples of Facial Action Units (AUs) defined by the FACS system [25,26]

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

Few methods have been reported for automatic AU detection in face image sequences [37,51,78]. Some researchers described patterns of facial motion that correspond to a few specific AUs, but did not report on actual recognition of these AUs (e.g. [10,29,42,43,76]). Only recently there has been an emergence of efforts toward explicit automatic analysis of facial expressions into elementary AUs [47,78]. For instance, the Machine Perception group at UCSD has proposed several methods for automatic AU coding of input face video. To detect 6 individual AUs in face image sequences free of head motions, Bartlett et al. [7] used a $61 \times 10 \times 6$ feed-forward neural network. They achieved 91% accuracy by feeding the pertinent network with the results of a hybrid system combining holistic spatial analysis and optical flow with local feature analysis. To recognize 8 individual AUs and 4 combinations of AUs in face image sequences free of head motions, Donato et al. [21] used Gabor wavelet representation and independent component analysis. They reported a 95.5% average recognition rate achieved by their method. The most recent work by Bartlett et al. [8] reports on accurate automatic recognition of 18 AUs (95% average recognition rate)

from near frontal-view face image sequences using Gabor wavelet features and a classification technique based on AdaBoost and Support Vector Machines (SVM). Another group that has focused on automatic FACS coding of face image sequences is the CMU group led by Takeo Kanade and Jeff Cohn. To recognize 8 individual AUs and 7 combinations of AUs in face image sequences free of head motions, Cohn et al. [16] used facial feature point tracking and discriminant function analysis and achieved an 85% average recognition rate. Tian et al. [77] used lip tracking, template matching and neural networks to recognize 16 AUs occurring alone or in combination in near frontal-view face image sequences. They reported an 87.9% average recognition rate.

Our group also reported on multiple efforts toward automatic analysis of facial expressions into atomic facial actions. The majority of this previous work concerns automatic AU recognition in static face images [50,53]. To our best knowledge, these systems are the first (and at this moment the only) to handle AU detection in static face images. However, these works are not relevant to the present goals, since they cannot handle video streams inherent in AmI applications. Only recently, our group has focused on automatic FACS coding of face video. To recognize 15 AUs occurring alone or in combination in near frontal-view face image sequences, Valstar et al. [80] used temporal templates (Section 3.2) and compared two classification techniques: (i) a combined k-Nearest-Neighbor and rule-based classifier, and (ii) a SNoW classifier. An average recognition rate ranging from 56% to 68% has been achieved. Except for this work, and based upon the tracked movements of facial characteristic points, we mainly experimented with rule-based [48,49] and SVM-based methods [81] for recognition of AUs in either near frontal-view (Fig. 10) or near profile-view (Fig. 11) face image sequences.

A basic understanding of how to achieve automatic AU detection from the profile view of the face is necessary if a technological framework for automatic AU detection from multiple views of the face is to be established [49]. The automatic AU detection from the profile view of the face was deemed the most promising method for achieving robust AU detection [86], independent of rigid head movements that can cause changes in the viewing angle and the visibility of the tracked face and its features. To our knowledge, our system for AU recognition from face profile-view image sequences is the first (and at this moment the only) to address this problem.

In contrast to the aforementioned methods developed elsewhere, which address mainly the problem of spatial modeling of facial expressions, the methods proposed by our group address the problem of temporal modeling of facial expressions as well. In other words, the methods proposed here are very suitable for encoding temporal activation patterns (onset \rightarrow apex \rightarrow offset) of AUs shown in an input face video. This is of importance for there is now a growing body of psychological research that argues that temporal dynamics of facial behavior (i.e., the timing and the duration of facial activity) is a critical factor for the interpretation of the observed behavior [67]. For example, Schmidt and Cohn [70] have shown that spontaneous smiles, in contrast to posed smiles, are fast in onset, can have multiple AU12 apexes (i.e., multiple rises of the mouth corners), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1 second. Since it takes more than

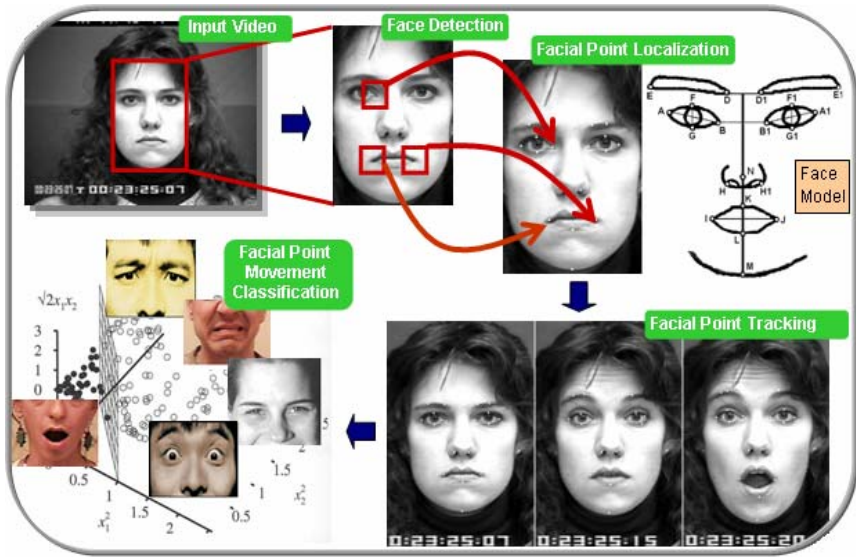


Fig. 10. Outline of our AU detectors from frontal-view face image sequences [48,81]

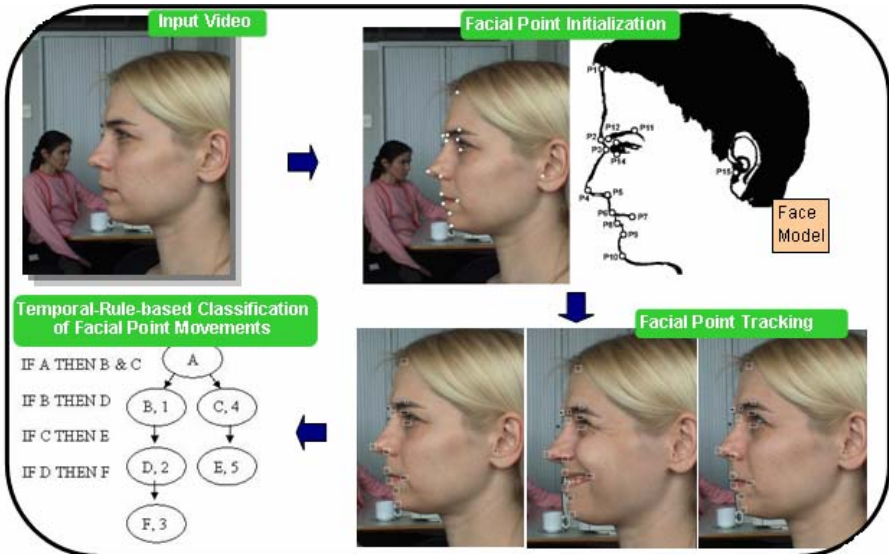


Fig. 11. Outline of our AU detector from profile-view face image sequences [49]

one hour to manually score 100 still images or a minute of videotape in terms of AUs and their temporal segments [25], it is obvious that automated tools for the detection of AUs and their temporal dynamics would be highly beneficial. To our best knowl-

edge, our systems are the first (and at this moment the only) to explicitly handle temporal segments of AUs.

To recognize a set of 27 AUs occurring alone or in combination in a near frontal-view face image sequence [48], we proceed under 2 assumptions (as defined for video samples of either the Cohn-Kanade [37] or the MMI facial expression database [56]): (1) the input image sequence is non-occluded near frontal-view of the face, and (2) the first frame shows a neutral expression and no head rotations. To handle possible in-image-plane head rotations and variations in scale of the observed face, we register each frame of the input image sequence with the first frame based on three referential points (Fig. 10): the tip of the nose (N) and the inner corners of the eyes (B and B1). We use these points as the referential points because of their stability with respect to non-rigid facial movements: facial muscle actions do not cause physical displacements of these points. Each frame is registered with the first frame by applying an affine transformation. Except of N, B and B1, which are tracked in unregistered input video sequences, other facial fiducial points are tracked in the registered input image sequence. Typical tracking results are shown in Fig. 9. Based upon the changes in the position of the fiducial points, we measure changes in facial expression. Changes in the position of the fiducial points are transformed first into a set of mid-level parameters for AU recognition. We defined two parameters: *up/down(P)* and *inc/dec(PP')*. Parameter *up/down(P)* = $y(P_{i1}) - y(P_i)$ describes upward and downward movements of point P and parameter *inc/dec(PP')* = $PP'_{i1} - PP'_i$ describes the increase or decrease of the distance between points P and P' . Based upon the temporal consistency of mid-level parameters, a rule-based method encodes temporal segments (onset, apex, offset) of 27 AUs occurring alone or in combination in the input face videos. For instance, to recognize the temporal segments of AU12 (Table 2), which pulls the mouth corners upwards in a smile, we exploit the following temporal rules (for experiments with a SVM-based binary classifier instead of rules, see [81]):

```

IF (up/down(I) >  $\epsilon$  AND inc/dec(NI) <  $-\epsilon$ )
  OR (up/down(J) >  $\epsilon$  AND inc/dec(NJ) <  $-\epsilon$ ) THEN AU12-p
IF AU12-p AND { (up/down(I)) $t$  > [up/down(I)] $t-1$  )
  OR ( [up/down(J)] $t$  > [up/down(J)] $t-1$  ) } THEN AU12-onset
IF AU12-p AND { ( | [up/down(I)] $t$  - [up/down(I)] $t-1$  | ≤  $\epsilon$  )
  OR ( | [up/down(J)] $t$  - [up/down(J)] $t-1$  | ≤  $\epsilon$  ) } THEN AU12-apex
IF AU12-p AND { (up/down(I)) $t$  < [up/down(I)] $t-1$  )
  OR ( [up/down(J)] $t$  < [up/down(J)] $t-1$  ) } THEN AU12-offset

```

Fig. 12 illustrates the meaning of these rules. The horizontal axis represents the time dimension (i.e., the frame number) and the vertical axis represents values that the mid-level feature parameters take. As implicitly suggested by the graphs of Fig. 12, I and/or J should move upward and be above their neutral-expression location to label a frame as an “AU12² onset”. The upward motion should terminate, resulting in a (relatively) stable temporal location of I and/or J, for a frame to be labeled as “AU12

² Since the upward motion of the mouth corners is the principle cue for the activation of AU12, the upward movement of the fiducial points I and/or J (i.e., point P7 in the case of the profile view of the face) is used as the criterion for detecting the onset of the AU12 activation. Reversal of this motion is used to detect the offset of this facial expression.

apex”. Eventually, I and/or J should move downward toward their neutral-expression location to label a frame as an “AU12 offset”. Note that, at the end of the offset phase, the graphs show a distinct increase of the values of the mid-level parameters, beyond their neutral-expression values. As shown by Schmidt and Cohn [70], this is typical for so-called “dampened” spontaneous smiles and in contrast to posed smiles.

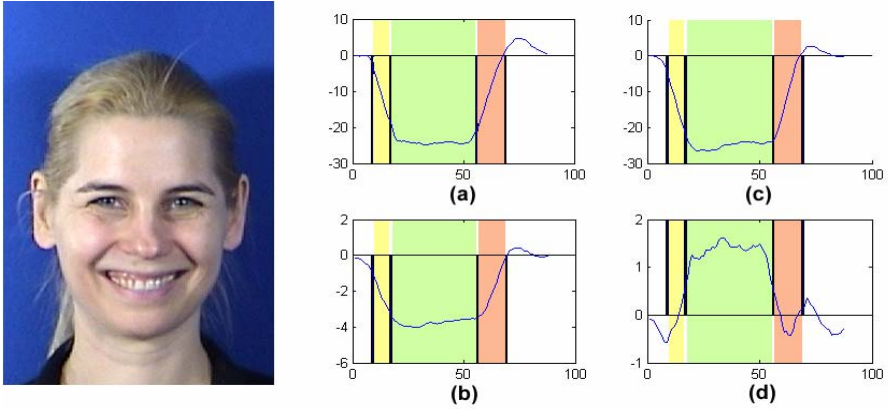


Fig. 12. The changes in x and y coordinates of points I and J (mouth corners) computed for 90 frames of an AU6+12+25 frontal-view face video (the apex frame is illustrated). (a)-(b) Point I. (c)-(d) Point J.

Similarly, to recognize a set of 27 AUs occurring alone, or in combination in a near profile-view face image sequence [49], we proceed under 2 assumptions (as defined for video samples of the MMI facial expression database [56]): (1) the input image sequence is non-occluded (left or right) near profile-view of the face with possible in-image-plane head rotations, and (2) the first frame shows a neutral expression. To make the processing robust to in-image-plane head rotations and translations as well as to small translations along the z-axis, we estimate a global affine transformation ϑ for each frame, and based on it we register the current frame to the first frame of the sequence. In order to estimate the global affine transformation, we track three referential points. These are (Fig. 11): the top of the forehead (P1), the tip of the nose (P4), and the ear canal entrance (P15). We use these points as the referential points because of their stability with respect to non-rigid facial movements. We estimate the global affine transformation ϑ as the one that minimizes the distance (in the least-squares sense) between the ϑ -based projection of the tracked locations of the referential points and these locations in the first frame of the sequence. The rest of the facial points illustrated in Fig. 11 are tracked in frames that have been compensated for the transformation ϑ . Typical tracking results are shown in Fig. 8. Changes in the position of the facial points are transformed first into a set of mid-level parameters for AU recognition described above. Based upon the temporal consistency of mid-level parameters, a rule-based method encodes temporal segments (onset, apex, offset) of 27 AUs occurring alone or in combination in the input face videos. For instance, to recognize the temporal segments of AU12, we exploit the following temporal rules:

```

IF ( $up/down(P7) > \epsilon$  AND  $inc/dec(P5P7) \geq \epsilon$ ) THEN AU12-p
IF AU12-p AND ( $([up/down(P7)]_t > [up/down(P7)]_{t-1})$ ) THEN
AU12-onset
IF AU12-p AND ( $(|[up/down(P7)]_t - [up/down(P7)]_{t-1}| \leq \epsilon)$ )
THEN AU12-apex
IF AU12-p AND ( $([up/down(P7)]_t < [up/down(P7)]_{t-1})$ ) THEN
AU12-offset

```

Fig. 13 illustrates the meaning of these rules. P7 should move upward, above its neutral-expression location, and the distance between points P5 and P7 should increase, exceeding its neutral-expression length, in order to label a frame as an “AU12 onset”. In order to label a frame as “AU12 apex”, the increase of the values of the relevant mid-level parameters should terminate. Once the values of these mid-level parameters begin to decrease, a frame can be labeled as “AU12 offset”. Note that the graphs of Fig. 13 show two distinct peaks in the increase of the pertinent mid-level parameters. According to [70], this is typical for spontaneous smiles.

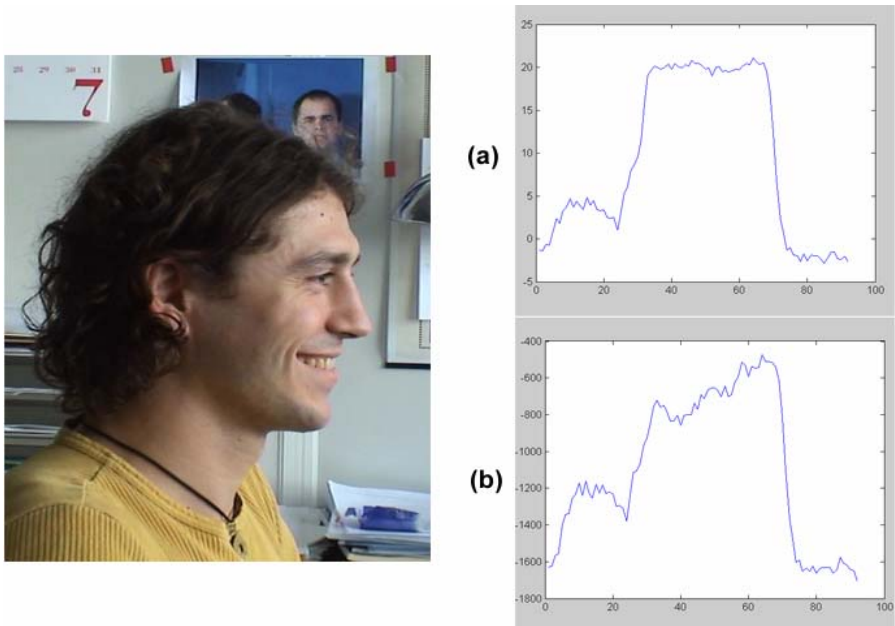


Fig. 13. The values of mid-level parameters (a) $up/down(P7)$ and (b) $inc/dec(P5P7)$ computed for 92 frames of AU6+12+25 face-profile video (the apex frame is illustrated)

We tested our method for AU coding in near frontal-view face image sequences on both Cohn-Kanade [37] and MMI facial expression database [56]. The accuracy of the method was measured with respect to the misclassification rate of each “expressive” segment of the input sequence, not with respect to each frame [48]. Overall, for 135 test samples from both databases, we achieved an average recognition rate of 90% sample-wise for 27 different AUs occurring alone or in combination in an input video.

Since Cohn-Kanade database does not contain images of faces in profile view (it contains only displays of emotions recorded in frontal facial view), the method for AU coding in near profile-view face video was tested on MMI facial expression database only. The accuracy of the method was measured with respect to the misclassification rate of each “expressive” segment of the input sequence[49]. Overall, for 96 test samples, we achieved an average recognition rate of 87% sample-wise for 27 different AUs occurring alone, or in combination, in an input video.

6 Facial Expression Interpretation and Facial Affect Recognition

As already noted above, virtually all systems for automatic facial expression analysis attempt to recognize a small set of universal/basic emotions [51,52]. However, pure expressions of “basic” emotions are seldom elicited; most of the time people show blends of emotional displays [38]. Hence, the classification of facial expressions into a single “basic”-emotion category is not realistic. Also, not all facial actions can be classified as a combination of the “basic” emotion categories. Think, for instance, about fatigue, frustration, anxiety, or boredom. In addition, it has been shown that the comprehension of a given emotion label and the ways of expressing the related affective state may differ from culture to culture and even from person to person [67]. Furthermore, not all facial actions should be associated with affective states. Think, for instance, about face-based interface-control systems for support of disabled users. Hence, pragmatic choices (user- and use-case-profiled choices) must be made regarding the selection of interpretation labels (such as affective labels) to be assigned by an automatic system to sensed facial signals. This is especially the case with AmI technologies, where one basic goal is to ensure that products are tailored to the user’s preferences, needs and abilities (Table 1).

We developed a case-based reasoning system that learns its expertise by having the user instruct the system on the desired (context-sensitive) interpretations of sensed facial signals [54]. To our best knowledge, it is the first (and at this moment the only) system facilitating user-profiled interpretation of facial expressions. We used it to achieve classification of AUs into the emotion categories learned from the user.

Since AUs can occur in more than 7000 combinations [69], the classification of AUs in an arbitrary number of emotion categories learned from the user is a very complex problem. To tackle this problem, one can apply either eager or lazy learning methods. Eager learning methods, such as neural networks, extract as much information as possible from training data and construct a general approximation of the target function. Lazy learning methods, such as case-based reasoning, simply store the presented data and generalizing beyond these data is postponed until an explicit request is made. When a query instance is encountered, similar related instances are retrieved from the memory and used to classify the new instance. Hence, lazy methods have the option of selecting a different local approximation of the target function for each presented query instance. Eager methods using the same hypothesis space are more restricted because they must choose their approximation before presented queries are observed. In turn, lazy methods are usually more appropriate for complex and incomplete problem domains than eager methods, which replace the training data with abstractions obtained by generalization and which, in turn, require an excessive amount

of training data. Therefore, we chose to achieve classification of the AUs detected in an input face video into the emotion categories learned from the user by case-based reasoning, a typical lazy learning method.

The utilized case base is a dynamic, incrementally self-organizing event-content-addressable memory that allows fact retrieval and evaluation of encountered events based upon the user preferences and the generalizations formed from prior input. Each event (case) is one or more micro-events, each of which is a set of AUs. Micro-events related by the goal of communicating one specific affective state are grouped within the same dynamic memory chunk. In other words, each memory chunk represents a specific emotion category and contains all micro-events to which the user assigned the emotion label in question. The indexes associated with each dynamic memory chunk comprise individual AUs and AU combinations that are most characteristic for the emotion category in question. Finally, the micro-events of each dynamic memory chunk are hierarchically ordered according to their typicality: the larger the number of times a given micro-event occurred, the higher its hierarchical position within the given chunk. The initial endowment of the dynamic memory is achieved by asking the user to associate an interpretation (emotion) label to a set of 40 typical facial expressions (micro-events that might be hardwired to emotions according to [46]). Fig. 14 illustrates a number of examples of the utilized stimulus material.



Fig. 14. Sample stimulus images from the MMI Facial Expression Database [56] used for initial endowment of the case base. Left to right: AU1+2, AU10, AU6+12, AU15+17.

The classification of the detected AUs into the emotion categories learned from the user is further accomplished by case-based reasoning about the content of the dynamic memory. To solve a new problem of classifying a set of input AUs into the user-defined interpretation categories, the following steps are taken:

1. Search the dynamic memory for similar cases, retrieve them, and interpret the input set of AUs using the interpretations suggested by the retrieved cases.
2. If the user is satisfied with the given interpretation, store the case in the dynamic memory. Otherwise, adapt the memory according to user-provided feedback on the interpretation he associates with the input facial expression.

The utilized retrieval and adaptation algorithms employ a pre-selection of cases that is based upon the clustered organization of the dynamic memory, the indexing structure

of the memory, and the hierarchical organization of cases within the clusters/ chunks according to their typicality [54].

Two validation studies on a prototype system have been carried out. The question addressed by the 1st validation study was: How acceptable are the interpretations given by the system after it is trained to recognize 6 basic emotions? The question addressed by the 2nd validation study was: How acceptable are the interpretations given by the system, after it is trained to recognize an arbitrary number of user-defined interpretation categories? In the first case, a human FACS coder was asked to train the system. In the second case, a lay expert, without formal training in emotion signals recognition, was asked to train the system. The same expert used to train the system was used to evaluate its performance, i.e., to judge the acceptability of interpretations returned by the system. For basic emotions, in 100% of test cases the expert approved of the interpretations generated by the system. For user-defined interpretation categories, in 83% of test cases the lay expert approved entirely of the interpretations and in 14% of test cases the expert approved of most but not of all the interpretation labels generated by the system for the pertinent cases. These findings indicate that the facial expression interpretation achieved by the system is rather accurate.

7 Conclusions

Automating the analysis of facial signals, especially rapid facial signals (i.e., AUs), is important to realize context-sensitive, face-based (multimodal) ambient interfaces, to advance studies on human emotion and affective computing, and to boost numerous applications in fields as diverse as security, medicine, and education. This paper provided an overview of the efforts of our research group in approaching this goal. To summarize:

- Our methods for automatic facial feature point detection and tracking extend the state of the art in facial point detection and tracking in several ways, including the number of facial points detected (20 in total), the accuracy of the achieved detection (93% of the automatically detected points were displaced in any direction, horizontal or vertical, less than 5% of the inter-ocular distance [83]), the accuracy, and the robustness of the tracking scheme including the invariance to noise, occlusion, clutter and changes in the illumination intensity (inherent in Particle Filtering with Factorized Likelihoods [57,58]).
- Our approaches to automatic AU coding of face image sequences extend the state of the art in the field in several ways, including the facial view (profile), the temporal segments of AUs (onset, apex, offset), the number (27 in total), and the difference in AUs (e.g. AU29, AU36) handled. To wit, the automated systems for AU detection from face video that have been reported so far do not deal with the profile view of the face, cannot handle temporal dynamics of AUs, cannot detect out-of-plane movements such as thrusting the jaw forward (AU29), and, at best, can detect 16 to 18 AUs (from in total 44 AUs). The basic insights in how to achieve automatic detection of AUs in profile-face videos and how to realize automatic detection of temporal segments of AUs in either frontal- or profile-view face image sequences can aid and abet further research on facial expression symmetry, spontaneous vs. posed facial expressions, and facial expression recognition from multiple facial views [48,49].

- We also proposed a new facial expression interpretation system that performs classification of AUs into the emotion categories learned from the user [54]. Given that the previously reported systems for high-abstraction-level analysis of facial expressions are able to classify facial expressions only in one of the 6 basic emotion categories, our facial expression interpreter extends the state of the art in the field by enabling facial signal interpretation in a user-adaptive manner. Further research on accomplishing context-sensitive (user-, task-, use-case-, environment-dependent) interpretation of facial (or any other behavioral) signals can be based upon our findings. This is especially important for AmI technologies where one basic goal is to ensure that products are tailored to the user's preferences, needs and abilities.

However, our methods cannot recognize the full range of facial behavior (i.e. all 44 AUs defined in FACS); they detect up to 27 AUs occurring alone or in combination in near frontal- or profile-view face image sequences. A way to deal with this problem is to look at diverse facial features. Although it has been reported that methods based on geometric features are usually outperformed by those based on appearance features using, e.g., Gabor wavelets or eigenfaces [7], our studies have shown that this claim does not always hold [49,81]. We believe, however, that further research efforts toward combining both approaches are necessary if the full range of human facial behavior is to be coded in an automatic way.

If we consider the state of the art in face detection and facial point localization and tracking, then noisy and partial data should be expected. As remarked by Pantic et al. [47,52], a facial expression analyzer should be able to deal with these imperfect data and to generate its conclusion so that the certainty associated with it varies with the certainty of face and facial point localization and tracking data. Our point tracker is very robust to noise, occlusion, clutter and changes in lighting conditions and it deals with inaccuracies in facial point tracking using a memory-based process that takes into account the dynamics of facial expressions [49,57,58]. However, our methods do not calculate the output data certainty by propagating the input data certainty (i.e. the certainty of facial point tracking). Future work on this issue aims at investigating the use of measures that can express the confidence to facial point tracking and that can facilitate both a more robust AU recognition and the assessment of the certainty of the performed AU recognition.

Finally, our methods assume that the input data are near frontal- or profile-view face image sequences showing facial displays that always begin with a neutral state. In reality, such assumption cannot be made; variations in the viewing angle should be expected. Also, human facial behavior is more complex and transitions from a facial display to another do not have to involve intermediate neutral states. Consequently, our facial expression analyzers cannot deal with spontaneously occurring (unposed) facial behavior. In turn, actual deployment of our methods in ambient interfaces and AmI sensing technologies is still in the relatively distant future. There are a number of related issues that should be addressed. How to achieve parsing of the stream of facial and head movements not under volitional control? What properties should automated analyzers of human expressive behavior have in order to be able to analyze human spontaneous behavior? How should one elicit spontaneous human expressive behavior, including genuine emotional responses, necessary for the training automated systems? How should the grammar of human expressive behavior be learned?

Tian et al. [78] and Pantic et al. [52,55] have discussed some of these aspects of automated facial expression analysis and they form the main focus of our current and future research efforts. Yet, since the complexity of these issues concerned with the interpretation of human behavior at a deeper level is tremendous and spans several different disciplines in computer and social sciences, we believe that a large, focused, interdisciplinary, international program directed towards computer understanding and responding to human behavioral patterns (as shown by means of facial expressions and other modes of social interaction) should be established if we are to experience breakthroughs in human-computer and ambient interface designs.

Acknowledgements

The work of M. Pantic is supported by the Netherlands Organization for Scientific Research Grant EW-639.021.202.

References

- [1] Aarts, E.: Ambient Intelligence – Visualizing the Future. Proc. Conf. Smart Objects & Ambient Intelligence (2005) (<http://www.soc-eusai2005.org/>)
- [2] Ambady, N., Rosenthal, R.: Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin*, Vol. 111, No. 2 (1992) 256-274
- [3] Anderson, K., McOwan, P.W.: A Real-Time Automated System for Recognition of Human Facial Expressions. *IEEE Trans. Systems, Man, and Cybernetics, Part B*. Vol. 36, No. 1 (2006) 96-105
- [4] Aristotle: *Physiognomonica*. In: Ross, W.D. (ed.): *The works of Aristotle*. Clarendon, Oxford (nd/1913) 805-813
- [5] Baker, S., Matthews, I., Xiao, J., Gross, R., Kanade, T.: Real-time non-rigid driver head tracking for driver mental state estimation. Proc. World Congress on Intelligent Transportation Systems (2004) (http://www.ri.cmu.edu/projects/project_448.html)
- [6] Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. *J. Computer Vision*, Vol. 12, No. 1 (1994) 43-78
- [7] Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Measuring facial expressions by computer image analysis. *Psychophysiology*, Vol. 36 (1999) 253-263
- [8] Bartlett, M.S., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.R.: Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions. Proc. Conf. Systems, Man, and Cybernetics, Vol. 1 (2004) 592-597.
- [9] Bassili, J.N.: Facial Motion in the Perception of Faces and of Emotional Expression. *J. Experimental Psychology*, Vol. 4 (1978) 373-379
- [10] Black, M., Yacoob, Y.: Recognizing facial expressions in image sequences using local parameterized models of image motion. *Computer Vision*, Vol. 25, No. 1 (1997) 23-48
- [11] Bobick, A.F., Davis, J.W.: The Recognition of Human Movement Using Temporal Templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3 (2001) 257-267
- [12] Bowyer, K.W.: Face Recognition Technology – Security vs. Privacy. *IEEE Technology and Society Magazine*, Vol. 23, No. 1 (2004) 9-19
- [13] Bruce, V.: *Recognizing Faces*. Lawrence Erlbaum Assoc., Hove (1986)

- [14] Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S.: Facial expression recognition from video sequences – temporal and static modeling. *Computer Vision and Image Understanding*, Vol. 91 (2003) 160-187
- [15] Cohn, J.F., Reed, L.I., Ambadar, Z., Xiao, J., Moriyma, T.: Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. *Proc. Conf. Systems, Man and Cybernetics*, Vol. 1 (2004) 610-616
- [16] Cohn, J.F., Zlochower, A.J., Lien, J., Kanade, T.: Automated face analysis by feature point tracking has high concurrent validity with manual faces coding, *Psychophysiology*, Vol. 36 (1999) 35-43
- [17] Cristinacce, D., Cootes, T.F.: A Comparison of Shape Constrained Facial Feature Detectors. *Proc. Conf. Automatic Face and Gesture Recognition (2004)* 375-380
- [18] Darwin, C.: *The expression of the emotions in man and animals*. University of Chicago Press, Chicago (1965 / 1872)
- [19] DeCarlo, D., Metaxas, D.: The integration of optical flow and deformable models with applications to human face shape and motion estimation. *Proc. Conf. Computer Vision and Pattern Recognition (1996)* 231-238
- [20] Dishman, E.: Inventing wellness systems for aging in place. *IEEE Computer Magazine, Spec. Issue on Computing and the Aging*, Vol. 37, No. 5 (2004) 34-41
- [21] Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Classifying Facial Actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10 (1999) 974-989
- [22] Ekman, P.: *Emotions Revealed*. Times Books, New York (2003)
- [23] Ekman, P., Friesen, W.V.: *The repertoire of nonverbal behavior*. *Semiotica*, Vol. 1 (1969) 49-98
- [24] Ekman, P., Friesen, W.V.: *Unmasking the face*. Prentice-Hall, New Jersey (1975)
- [25] Ekman, P., Friesen, W.V.: *Facial Action Coding System*. Consulting Psychologist Press, Palo Alto (1978)
- [26] Ekman, P., Friesen, W.V., Hager, J.C.: *Facial Action Coding System. A Human Face*, Salt Lake City (2002)
- [27] Fasel, I., Fortenberry, B., Movellan, J.R.: GBoost: A generative framework for boosting with applications to real-time eye coding. *Computer Vision and Image Understanding*, under review (<http://mplab.ucsd.edu/publications/>)
- [28] Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, Vol. 28, No. 2 (2000) 337-374
- [29] Gokturk, S.B., Bouguet, J.Y., Tomasi, C., Girod, B.: Model-based face tracking for view-independent facial expression recognition. *Proc. Conf. Automatic Face and Gesture Recognition (2002)* 272-278.
- [30] Gross, T.: Ambient Interfaces – Design Challenges and Recommendations. *Proc. Conf. Human-Computer Interaction (2003)* 68-72
- [31] Gu, H., Ji, Q.: Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications*, Vol. 16, No. 2 (2005) 105-115
- [32] Guo, G., Dyer, C.R.: Learning From Examples in the Small Sample Case – Face Expression Recognition. *IEEE Trans. Systems, Man, and Cybernetics, Part B*. Vol. 35, No. 3 (2005) 477-488
- [33] Haykin, S., de Freitas, N. (eds.): *Special Issue on Sequential State Estimation*. *Proceedings of the IEEE*, vol. 92, No. 3 (2004) 399-574
- [34] Isard, M., Blake, A.: Condensation – conditional density propagation for visual tracking. *J. Computer Vision*, Vol. 29, No. 1 (1998) 5-28

- [35] Jacobs, D.W., Osadchy, M., Lindenbaum, M.: What Makes Gabor Jets Illumination In-sensitive? (<http://rita.osadchy.net/papers/gabor-3.pdf>)
- [36] Kalman, R.E.: A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.*, Vol. 82 (1960) 35-45
- [37] Kanade, T., Cohn, J., Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proc. Conf. Automatic Face and Gesture Recognition (2000)* 46-53.
- [38] Keltner, D., Ekman, P.: Facial Expression of Emotion. In: Lewis, M., Haviland-Jones, J.M. (eds.): *Handbook of Emotions*. 2nd edition. The Guilford Press, New York (2004) 236-249
- [39] Li, S.Z., Jain, A.K. (eds.): *Handbook of Face Recognition*. Springer, New York (2005)
- [40] van Loenen, E.J.: On the role of Graspable Objects in the Ambient Intelligence Paradigm. *Proc. Conf. Smart Objects (2003)* (<http://www.grenoble-soc.com/>)
- [41] Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. *Proc. Conf. Artificial Intelligence (1981)* 674-679
- [42] Martinez, A.M.: Matching expression variant faces. *Vision Research*, Vol. 43 (2003) 1047-1060
- [43] Mase, K.: Recognition of facial expression from optical flow. *IEICE Transactions*, Vol. E74, No. 10 (1991) 3474-3483
- [44] Moghaddam, B., Pentland, A.: Probabilistic Visual Learning for Object Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7 (1997) 696-710
- [45] Norman, D.A.: *The Invisible Computer*. MIT Press, Cambridge (1999)
- [46] Ortony, A., Turner, T.J.: What is basic about basic emotions? *Psychological Review*, Vol. 74 (1990) 315-341
- [47] Pantic, M.: Face for Interface. In: Pagani, M. (ed.): *The Encyclopedia of Multimedia Technology and Networking 1*. Idea Group Reference, Hershy (2005) 308-314
- [48] Pantic, M., Patras, I.: Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. *Proc. Conf. Systems, Man, and Cybernetics (2005)*
- [49] Pantic, M., Patras, I.: Dynamics of Facial Expressions – Recognition of Facial Actions and their Temporal Segments from Face Profile Image Sequences. *IEEE Trans. Systems, Man, and Cybernetics, Part B*. Vol. 36 (2006)
- [50] Pantic, M., Rothkrantz, L.J.M.: Expert system for automatic analysis of facial expression. *Image and Vision Computing*, Vol. 18, No. 11 (2000) 881-905
- [51] Pantic, M., Rothkrantz, L.J.M.: Automatic Analysis of Facial Expressions – The State of the Art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12 (2000) 1424-1445
- [52] Pantic, M., Rothkrantz, L.J.M.: Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE, Spec. Issue on Human-Computer Multimodal Interface*, Vol. 91, No. 9 (2003) 1370-1390
- [53] Pantic, M., Rothkrantz, L.J.M.: Facial Action Recognition for Facial Expression Analysis from Static Face Images, *IEEE Trans. Systems, Man, and Cybernetics, Part B*. Vol. 34, No. 3 (2004) 1449-1461
- [54] Pantic, M., Rothkrantz, L.J.M.: Case-based reasoning for user-profiled recognition of emotions from face images, *Proc. Conf. Multimedia and Expo*, Vol. 1 (2005) 391-394
- [55] Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective Multimodal Human-Computer Interaction, *Proc. ACM Conf. Multimedia (2005)*
- [56] Pantic, M., Valstar, M.F., Rademaker, R., Maat, L.: Web-based database for facial expression analysis, *Proc. Conf. Multimedia and Expo (2005)* (<http://www.mmifacedb.com/>)

- [57] Patras, I., Pantic, M.: Particle Filtering with Factorized Likelihoods for Tracking Facial Features. Proc. Conf. Automatic Face and Gesture Recognition (2004) 97-102
- [58] Patras, I., Pantic, M.: Tracking Deformable Motion. Proc. Conf. Systems, Man, and Cybernetics (2005)
- [59] Pentland, A.: Looking at people – Sensing for ubiquitous and wearable computing. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No. 1 (2000) 107-119
- [60] Pentland, A., Moghaddam, B., Starner, T.: View-Based and Modular Eigenspaces for Face Recognition. Proc. Conf. Computer Vision and Pattern Recognition (1994) 84-91
- [61] Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
- [62] Pitt, M.K., Shephard, N.: Filtering via simulation: auxiliary particle filtering. J. Amer. Stat. Assoc., Vol. 94 (1999) 590-599
- [63] Preece, J., Rogers, Y., Sharp, H.: Interaction Design – Beyond Human-Computer Interaction. John Wiley & Sons, New York (2002)
- [64] Raisinghani, M.S., Benoit, A., Ding, J., Gomez, M., Gupta, K., Gusila, V., Power, D., Schmedding, O.: Ambient Intelligence – Changing Forms of Human-Computer Interaction and their Social Implications. J. Digital Information, Vol. 5, No. 4 (2004) 1-8
- [65] Remagnino, P., Foresti, G.L.: Ambient Intelligence – A New Multidisciplinary Paradigm. IEEE Trans. Systems, Man, and Cybernetics, Part A, Spec. Issue on Ambient Intelligence, Vol. 35, No. 1 (2005) 1-6
- [66] Rowley, H., Baluja, S., Kanade, T.: Neural Network-Based Face Detection. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, No. 1 (1998) 23-38
- [67] Russell, J.A., Fernandez-Dols, J.M. (eds.): The Psychology of Facial Expression. Cambridge University Press, Cambridge (1997)
- [68] Samal, A., Iyengar, P.A.: Automatic recognition and analysis of human faces and facial expressions: A survey. Pattern Recognition, Vol. 25, No. 1 (1992) 65-77
- [69] Scherer, K.R., Ekman, P. (eds.): Handbook of methods in non-verbal behavior research. Cambridge University Press, Cambridge (1982)
- [70] Schmidt, K.L., Cohn, J.F.: Dynamics of facial expression: Normative characteristics and individual differences. Proc. Conf. Multimedia and Expo (2001) 547-550
- [71] Shadbolt, N.: Ambient Intelligence. IEEE Intelligent Systems, Vol. 18, No. 4 (2003) 2-3
- [72] Shi, J., Tomasi, C.: Good features to track. Proc. Conf. Computer Vision and Pattern Recognition (1994) 593-600
- [73] Stephanidis, C., Akoumianakis, D., Sfyarakis, M., Paramythis, A.: Universal accessibility in HCI. Proc. ERCIM Workshop. User Interfaces For All (1998) (<http://ui4all.ics.forth.gr/UI4ALL-98/proceedings.html>)
- [74] Streitz, N., Nixon, P.: The Disappearing Computer. ACM Communications, Spec. Issue on The Disappearing Computer, Vol. 48, No. 3 (2005) 33-35
- [75] Sung, K.K., Poggio, T.: Example-Based Learning for View-Based Human Face Detection. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, No. 1 (1998) 39-51
- [76] Tao, H., Huang, T.S.: Connected vibrations – a model analysis approach to non-rigid motion tracking. Proc. Conf. Computer Vision and Pattern Recognition (1998) 735-740
- [77] Tian, Y., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Analysis & Machine Intelligence, Vol. 23, No. 2 (2001) 97-115
- [78] Tian, Y.L., Kanade, T., Cohn, J.F.: Facial Expression Analysis. In: Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition. Springer, New York (2005)
- [79] Tscheligi, M.: Ambient Intelligence – The Next Generation of User Centeredness. ACM Interactions, Spec. Issue on Ambient Intelligence, Vol. 12, No. 4 (2005) 20-21
- [80] Valstar, M., Pantic, M., Patras, I.: Motion History for Facial Action Detection from Face Video. Proc. Conf. Systems, Man and Cybernetics, Vol. 1 (2004) 635-640

- [81] Valstar, M., Patras, I., Pantic, M.: Facial Action Unit Detection using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data. Proc. Conf. Computer Vision and Pattern Recognition (2005)
- [82] Viola, P., Jones, M.: Robust real-time object detection. Proc. Int'l Conf. Computer Vision, Workshop on Statistical and Computation Theories of Vision (2001)
- [83] Vukadinovic, D., Pantic, M.: Fully automatic facial feature point detection using Gabor feature based boosted classifiers. Proc. Conf. Systems, Man and Cybernetics (2005)
- [84] Weiser, M.: The world is not a desktop. ACM Interactions, Vol. 1, No. 1 (1994) 7-8
- [85] Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time Combined 2D+3D Active Appearance Models. Proc. Conf. Computer Vision and Pattern Recognition, Vol. 2 (2004) 535-542
- [86] Yacoob, Y., Davis, L., Black, M., Gavrila, D., Horprasert, T., Morimoto, C.: Looking at People in Action. In: Cipolla, R., Pentland, A. (eds.): Computer Vision for Human-Machine Interaction. Cambridge University Press, Cambridge (1998) 171-187
- [87] Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting faces in images: A survey. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, No. 1 (2002) 34-58
- [88] Zhai, S., Bellotti, V.: Introduction to Sensing-Based Interaction. ACM Trans. Computer-Human Interaction, Spec. Issue on Sensing-Based Interaction, Vol. 12, No. 1 (2005) 1-2
- [89] Zhang, Y., Ji, Q.: Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequence. IEEE Trans. Pattern Analysis & Machine Intelligence, Vol. 27, No. 5 (2005) 699-714
- [90] Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J.: Face Recognition – A literature survey. ACM Computing Surveys, Vol. 35, No. 4 (2003) 399-458