IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

# Facial Expression Analysis and Recognition: A Survey

Beat Fasel [1] and Jürgen Lüttin [2]

IDIAP–RR 99-19

December 1999

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

1 Beat.Fasel@idiap.ch
2 Juergen.Luettin@idiap.ch

IDIAP Research Report 99-19

# Facial Expression Analysis and Recognition: A Survey

Beat Fasel and Jürgen Lüttin

December 1999

**Abstract.** In recent years, facial expression analysis has become an active research area. Various approaches were made towards robust facial expression recognition, applying different image acquisition, analysis and classification methods. Facial expression analysis is an inherently multi-disciplinary field and it is important to look at it from all domains involved in order to gain insight on how to build reliable automated facial expression analysis systems. This fact has often been neglected in various implementations presented in the literature. Facial expressions reflect not only emotions but also cognitive processes, social interaction and physiological signals. Image analysis methods should be chosen carefully to allow for a maximum of information to be extracted from faces, in order that an appropriate classification stage is able to code also subtle facial changes that occur frequently in everyday situations. In this survey we compare the most prominent automated facial expression analysis methods and systems. We found those systems to deliver best results, that implement a combination of holistic and local image analysis and classify facial expressions with the FACS coding scheme. Of great interest are multi-modal systems that do not only rely on facial expression analysis, but integrate also emotional voice analysis for more reliable interpretation of facial expressions. Yet another promising approach to facial expression analysis are unified frameworks, where multiple face processing tasks complement each other and thus allow for synergy effects.

**Keywords.** Facial expression analysis – Facial expression recognition – Emotion recognition – Facial motion analysis – Face detection – Temporal modeling – Pose estimation – Motion estimation – Parametric models – Appearance-based models – FACS – EMFACS – MAX – Human-computer interaction – Multi-modal systems

# Contents

# 1 Introduction

Facial expression analysis goes well back into the last century. *Darwin* [115] demonstrated already in 1872 the universality of expressions and their continuity in man and animals. Earlier in this century, *Schlosberg* [135] established a basis for describing emotions using three dimensions: attention-rejection (A-R), pleasantness-unpleasantness (P-U) and an activation level. For example, emotions such as contempt and disgust have a high value of rejection while anger has a high degree of unpleasantness. In the seventies, *Ekman and Friesen* [122] have identified six primary emotions: happiness, sadness, fear, disgust, surprise and anger. Apart from these, they have also systematically cataloged thousands of different facial expressions. Facial expression analysis was primarily a research subject for psychologists but already in 1978, *Suwa et al.* [95] presented a preliminary investigation on automatic facial expression analysis from an image sequence.

In the nineties, facial expression research done in the field of computer science gained much inertia starting with the pioneering work of *Mase* [71]. The reasons for this renewed interest in facial expressions are multiple. On one hand, automatic facial expression analysis could only lift off due to the advancements accomplished in related research domains such as object detection, object recognition and object tracking as well as their direct application in face processing, such as face detection, recognition and tracking. On the other hand, facial expression analysis is a computationally intensive task and only recent developments in microprocessors allow to accomplish this task. Further on, image acquisition devices such as digital cameras are now produced for the mass marked at low prices. Applications using facial expression analysis can be envisaged in the near future, fostering further the interest of doing research in the domain of facial expression processing, be it facial expression recognition or facial expression animation.

*Emotion recognition* is not the same as *facial expression recognition.* Facial expressions are what we observe on faces and encompass various different displays of mental states and serve as a means for non-verbal communication. Emotions represent only one category of these displays and their interpretation demands the inclusion of context information and if possible audio support. Emotions can sometimes be interfered from facial expressions but direct mapping of facial expressions to emotional states can be difficult especially with feigned expressions.

## 1.1 Motivation and Applications

Using non-verbal information provided by facial expressions is of great importance in many areas including image understanding, speech recognition, image compression and synthetic face animation. In this survey we try to give an overview of the research in facial expression analysis that has been done in recent years and concentrate on *automatic facial expression analysis systems* that will be of interest in various circumstances. Current applications encompass:

- **Expression animation:** Simple *facial animations* of cartoon characters can be controlled by tracking marker points placed in the face of an actor. This allows to substantially reduce production cost of cartoons (e.g. *famous*[1], facial animation solutions, *x-ist*[2] realtime technologies)
- **Psychology:** Psychologists use *automated facial expression recognition systems* based on FACS (Facial Action Coding System) to classify facial expressions sequences of faces that were marked prior to recording on video [138]. Coded facial actions allow an exact determination of the appearance of facial expressions and leverages the task of mapping these to emotions.
- **Surgery:** Surgeons use automated systems to determine the well functioning of facial muscles after surgery, for a review see [150].

These approaches cover facial expressions processing with limited potential of interpretation. At present, reliable systems are mostly based on relatively simple marker tracking approaches. Marker

[1] http://www.famoustech.com/
[2] http://www.x-ist.de/

independent systems would allow for more natural setups, e.g. in psychology facial expression analysis could be achieved without the knowledge of the subject, therefore interference would be minimized. In the near future we will probably see a multitude of new application fields for facial expression analysis systems:

- **Computer Interfaces** should be alive and engaging rather than static displays that wait to be clicked on. An *advanced Human-Machine interface* displaying an animated face (inter*face*!) which reacts to human facial expression changes combined with speech recognition would facilitate communication with a computer much in the way a natural conversation between two human beings takes place. This may be achieved with audio-visual speech synthesis and *talking heads*. An advanced multi-modal[3] interface could exploit different verbal and non-verbal communication channels, such as gesture and facial expressions. The life of physically handicapped people could be facilitated by allowing them to use computers controlled by non-verbal communication, e.g. guidance of a cursor via gaze direction.
- **Communication over low bandwidth channels:** The upcoming *MPEG-4-SNHC* standard introduces the coding and synthesis of faces, where facial expressions are represented by a few parameters prior to transmission. This allows for very high compression rates[4]. Applications include *video phones, virtual conferences*, *personalized agent interfaces*, *email readers, avatars for games* and *chat rooms.*
- **Video indexing and retrieval:** Faces play an important role in human communication and are therefore often depicted, e.g. in films. Image databases tend to grow to enormous sizes and therefore powerful search tools are needed for quick access. Searching for images that contain faces in specific situations could be improved by systems that would allow for the distinction of facial expressions.
- **Film industry:** Computer animated photo-realistic synthetic faces guided by facial expressions of actors would allow for new effects and reduced risk in dangerous film situations.

We can distinguish *online* applications such as face aware computer interfaces or real-time animated synthetic faces. Video indexing and psychological analysis on the other hand can be done *offline.* **Virtual reality** will be of great importance in a more distant future, where either virtual lifelike autonomous agents or other peoples avatar will be able to get in touch with our personal avatars. Facial expressions would facilitate communication considerably. Since communication and storage capacities are limited, virtual bodies or faces need to be represented and animated with a minimum number of parameters, extracted by using appropriate facial expression analysis. Still another form of autonomous agents would be real-world representatives or humanoid robots, that are able to interact with human beings reading and displaying facial expressions.

## 1.2 Overview of the Next Sections

In the following chapters, we will first describe the multi-disciplinary nature of facial expressions, discuss methods and techniques used in automatic facial expression analysis systems and then review and the most important systems found in the literature. Comparisons of some of these systems will reveal strong and weak points as well as operation constraints and robustness towards environmental changes. Finally, the survey is concluded by a general discussion, where an outlook is given on possible future developments.

[3] Multimodal interfaces feature multiple channels that act on multiple modalities. A channel is a communication medium associated with a particular encoding method. Examples are the auditory channel that carries speech and the visual channel that carries facial expressions [96].

[4] http://www.research.att.com/˜osterman/AnimatedHead/index.html

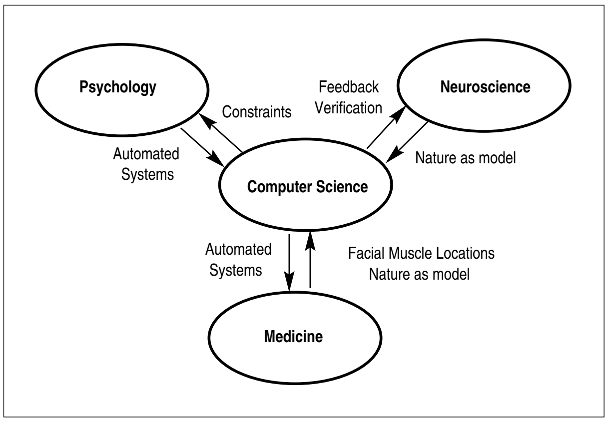


Figure 1: Relation of computer science with regard to the other research fields involved with facial expression analysis.

## 2 A Multi-disciplinary Field

Facial expression analysis is a typical multi-disciplinary field. It involves psychology, neurology, medicine and computer science. Psychologists are mainly interested in how facial expressions are related to emotions and thoughts, the context they occur in and how they can be interpreted. Facial expressions are of great interest because they reveal much about ongoing emotions. Neuroscientists and especially neurologists on the other hand try to gain information about the human brain itself, how it works and also how it communicates via non-verbal channels such as facial expressions. Computer scientists finally implement automated facial expression analysis systems and may benefit to a great deal from the other domains involved. Psychologists have defined accurate coding and categorization dictionaries for facial expressions. These are of great importance also for automated facial expression analysis systems. In addition, psychologist can at least show partially how facial expressions are related to emotions and thoughts and how these act as additional communication channels. Medicine provides detailed information on the arrangements of facial muscles. This is important for the construction of physical face models used abundantly in conjunction with animated synthetic heads. Facial expression analysis systems are of interest in the medical domain because these allow surgeons to verify the well functioning of facial muscles and nerves. Psychologists on the other hand use automated facial expression analysis systems to circumvent the time consuming task of manually coding facial expressions. Neurologists may verify models of how the brain works by gaining feedback from artificial systems built on similar principals using related techniques such as neural networks. Figure 1 shows how computer science is related to other other research fields involved with facial expression analysis. While the sciences of psychology, medicine and neuroscience are all centered on living beings, computer science has nothing much in common. Nonetheless, neurologists insight on the functioning of the brain has contributed substantially to the development of artificial systems that imitate its behavior. Biological systems can be a rich source for inspiration on how to implement systems using related methods such as for example retinally inspired vision acquisition methods.

### 2.1 Psychology

Psychologists are interested in mapping facial expressions to emotions and vice versa. Facial expressions do not only occur with emotions but display also cognitive processes and serve as a means of non-verbal communication independent of emotional display.

### 2.1.1 What the Face Reveals

According to psychoevolutionist theories, facial expressions play an important role in the communication of emotion and in the regulation of social interactions. The information provided by facial displays allows protagonists involved in social interactions to mutually appraise their emotional states and adjust their behavior in an appropriate way. One phenomenon that contributes to the complexity of emotion communication is the control that human beings have on their facial expressions. Emotions can be masked to some extent (poker face) and negative emotional display may be suppressed altogether in this manner. Often, controlled emotion is revealed by the absence of displayed emotion in a certain context where an emotional display would be normal. Emotions can be classified into *positive* like happiness, pride, relief and *negative* ones such as anger/irritation, anxiety/fear, sadness/disappointment and embarrassment/shame. In the regulative function of controlling facial expressions, smiles play an important role. Controlled smiles are often asymmetric smiles or smiles that occur with signs of negative emotions [131]. Facial expressions carry various signals apart from basic emotions, namely:

- **Illustrators** such as raising eyebrows while explaining a fact. These support the message given by a talking person.
- **Regulators** mark changes between a listener and a talking person. For example a listener gets bored by the subject of conversation and shows this accordingly by facial expressions.
- **Manipulators**, self-manipulative associated movements such as lip biting.
- **Emblems**, culture specific symbolic communicators such as winks.
- **Implicit feedback** such as nodding during an explanation provide a back-channel to a talking person. A listener shows that he has the same level of knowledge, respectively does not understand.
- **Relation signals** such as "we are still friends" (e.g. smile) can occur during conversations where opponents criticize each other.
- **Cognitive activity**, where mental processes such as concentration of reflection are displayed.

Emotion display is dependent on the context a given subject is in, e.g. a conversational situation where non-visual cues are important, further its cultural background, gender and approximate age[5]. Most facial displays can be superimposed by a basic *current emotional state* which leads to combinations of emotions displayed at the same time. Figure 2 shows mental activities that can elicit facial expression. Even though emotions are probably the main cause for facial expressions, we should preferably use *mental activities* when referring to the source of facial expressions instead of emotion alone. Apart from emotion, cognitive and psycho-physiological body signals such as sleepiness and pain as well as social interaction can be at the origin of facial expressions.

### 2.1.2 Other Non-verbal Communication Sources

The face is by no means the only non-verbal communication source, even though it is probably the most important one. Facial expressions or facial gestures are often accompanied by other *body gestures* such as head and arm movements that convey in general fewer information on a given situation, but may be at least temporally of ubiquitous importance. Apart from these visual cues, voice is important as it carries information in its intonation, c'est le ton qui fait la musique! Emotional display occurs often simultaneously in voice and on the face. This might lead to false interpretations when observing only one communication channel such as facial expressions. Consider for example the case where a subject displays a smile and says something accompanied by a menacing timber in the voice. Taking into

[5] Note that these features can be detected for both familiar and unfamiliar faces.

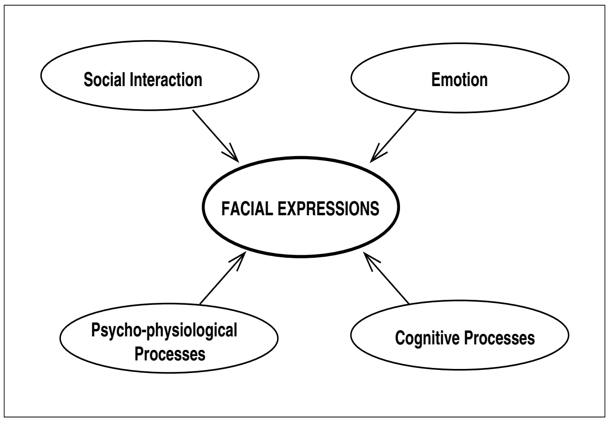


Figure 2: Mental and physical activities that can elicit facial expressions.

account both perceptual clues vision and acoustic as well as reasoning on how a given subject would respond in a certain situation, is mandatory for the interpretation of facial expressions. Contextual information is important: people expect different facial displays in a poker game, where bluffing is the norm than during a marriage proposal, where sincerity is expected.

#### 2.1.3 Some Aspects of Emotion and Facial Expression

Interesting is the function of emotion with respect to the following two interrelated regulatory processes [129][130]:

- The **intra-individual regulation** of thoughts and behavior. Major issues concern the intra-individual processes and the relation between emotion and cognition, particularly emotion-antecedent appraisal, as well as interactions between different emotion components, in particular the role of subjective experience or feeling state.
- The **inter-individual regulation** in social interactions. This process is studied in terms of how emotions and emotion expression regulate interpersonal interactions and, vice versa, how emotions are regulated during interactions.

In this framework, facial expressions can be seen as an interface between these two regulatory processes or systems. *Baenninger-Huber et al.* [111] have developed an epistemological model of the affective regulation process which contains four hierarchical levels, two levels of conceptual analysis and two levels of data analysis:

- **High-level concepts:** Top level models of emotional processes in relationship situations. An examples from social psychology are Mehrabians dimensions as like - dislike, potency and responsiveness. Models of affective regulation processes include concepts such as self-regulation, relationship-regulation, affective relatedness, resonance, response or dialogue regulation.
- **Low-level concepts:** Models that cover different definitions of emotion and models of the association between emotions, cognitions, psycho-physiological variables and expressive behavior. Examples are basic emotions [117], communicative affects versus self-feelings and state versus trigger affects.
- **High-level data:** Directly observable phenomena such as facial expression of basic emotions and their blends (FACS, EMFACS, MAX emotion interpretations), facial and bodily signs of complex emotions, such as shame and embarrassment, laughter, types of smile, turning away etc.

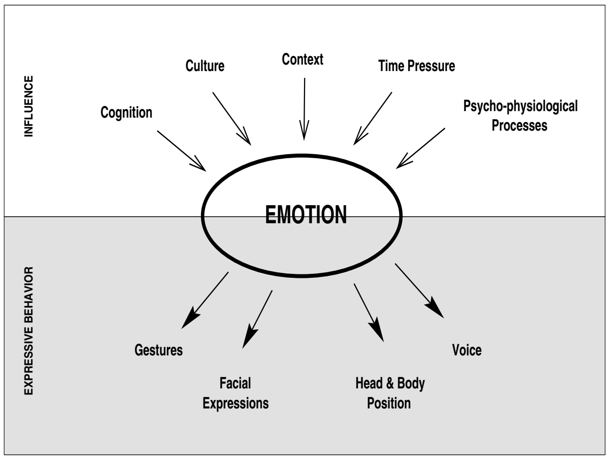


Figure 3: *Emotion and its interaction with the environment.* Note that facial expressions are *only one way* of displaying emotion, cognitive and psycho-physiological processes. Emotional voice, body gestures and positions are other important non-verbal channels.

- **Low-level data:** Objectively assessed non-verbal and verbal data, e.g. AUs, AU-combinations (see FACS further below), blinks, head positions, gaze direction, body movements, adaptors etc.

Low-level and high-level data are linked by operationalizations which were empirically tested. Working hypotheses may connect high-level data with the above listed theoretical concepts.

#### 2.1.4 Facial Expression Recognition versus Emotion Recognition

At this point we would like to state the difference between facial expression recognition and emotion recognition. *Facial expression recognition* deals with the classification of facial actions with the aid of a facial expressions dictionary, possibly containing all perceptible changes that may occur on a face. Once facial actions are clearly labeled, they may be interpreted. That's what we call *emotion recognition*. It is possible of course to directly map facial actions into mental activities. Emotions are not only displayed via facial expressions, but also via body gestures. Furthermore, emotion recognition often demands *understanding* of a given situation with full context information. Clearly, we should not confuse emotion recognition with facial expression recognition. The latter deals with the recognition of facial actions, whereas the former encompasses the full range of expressive behavior, as well as other influences such as context information, time pressure and cultural background. Figure3 displays emotion and its interaction with the environment and illustrates the fact that emotions cannot be deduced solely by interpreting facial expressions.

### 2.2 Neuroscience

How nature achieves the face processing tasks is of general interest as we may draw inspiration and learn from biological solutions when designing machines that face the same problems [143]. Neuroscientists are interested in facial expressions, as these provide a window to the brain. Of interest is how the human brain analysis facial expressions and how it interprets them. It is probably conditioned for faces even before birth.[6] In oder to represent the vast information that comes from visual stimulus, the brain creates models, especially interesting for us when it comes to faces. Human brains seem

[6]Interesting to note is that horizontal face features make a face look like a face which can easily be observed when looking at an image of a face that has been gradient filtered for vertical lines [21]. Of interest is also the fact that features in a 45 degree angle (mouth, eyebrows, edges in general) are of importance and serve as distincitive natural markers that can be used for facial expressions recognition.

to compress and memorize important objects with models, thus allowing for representations with a minimum number of parameters and still allow flexibility with changes. For example we are able to recognize faces of people that we haven't seen for a long time. Even though their appearance has changed due to aging, different hair-style, make-up, glasses or hair-covering, we can still quite reliably recognize them. Apparently the brain represents this information using population codes [145].

Facial expression recognition depends on several factors, e.g. familiarity with the face of a given person, cultural experience with different types of expression displays, attention given to the face, as well as non-visual cues. It is generally believed to be an early perceptual process and is separated from those involved in feature detection [144]. Furthermore, animal communication through facial expressions is often thought to require very high-level, almost cognitive processing, yet pigeons can be trained to recognize faces and babies can do so as early as 3 hours after birth [146]. Statistical analysis of neurons in the temporal cortex suggest, that they may be coding the physical dimensions of faces, whereas the responses of neurons in the amygdala may be better characterized by the social significance or emotional value of the face [148]. Others areas of the visual cortex are involved in representing facial expressions, lesions in some areas can lead to deficiencies in facial expression recognition. This leads to the assumption that facial expression analysis takes place in specialized regions, separated from other face processing tasks such as face recognition. On the other hand, facial display on a known face may be interpreted with more accuracy. Finally, it seems that humans analyze facial expressions both locally, focusing on facial features as well as holistically, perceiving the whole face at once.

*Bruce* [140] and *Cabeza et al.* [142] found that if participants studied faces shown at an angle plus or minus 30 degrees from the neutral position, they did later not tend to find the neutral face position as familiar as the studied ones. These findings suggest that different exemplars of a same person's face may be amalgamated in memory in a viewpoint-specific way, with face representations established separately at a probably small number of discrete angels. This is of interest, as facial expressions could be studied in conjunction with these face prototypes, see also the *complete facial expression analysis framework* section further below.

## 2.3 Medicine

*Surgeons* are interested in measuring the capability of patients to produce facial expression before and after face surgery. This may be necessary after a facial neuromuscular disfunction, an impairment of the facial neuromuscular motor system. The evaluation of the facial motor nerve function requires that movements of the facial musculature be elicited, either by a verbal command or by an external stimulation applying electrical current. Measures of the amplitude of facial expressions as well as symmetry may be important characteristics. An independent line of research has resulted in the development of various facial nerve grading systems [151][152][154], where facial movements are evoked by voluntary contracture. See also [150] for a review on methods for the objective topographic evaluation of facial nerve functions.

In *neurology*, analysis of facial expressions may help to reveal the location and type of brain lesions. For example Parkinson's decease, a disorder of the dopaminerfic system, is associated with amimia, or reduction of spontaneous facial activity [149], including a decreased eye blink rate. Subcortical lesions may also lead to increased facial activity, as in Meige's disease or Brueghel's syndrome, thought to result from a disorder of the basal ganglia [124].

In the *psychiatric* domain, one has observed that schizophrenia and psychosomatic illness lead to blunting of expressions, both in patients and in control subjects talking to these patients, who did not consciously know, they were talking to mentally ill people [153][155].

## 2.4 Computer Science

Computer scientist may benefit from the advances in research that were made in the last few years both in psychology and neural sciences. Among these we may cite the introduction of the facial

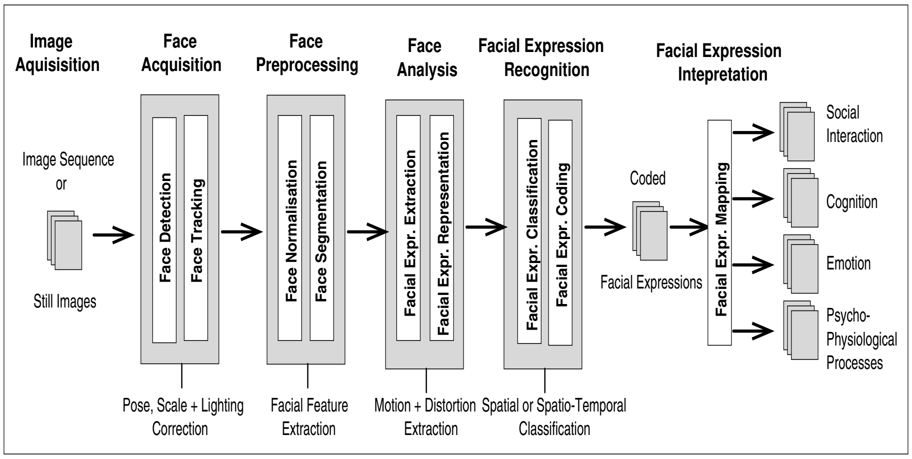


Figure 4: Generic Facial Expression Analysis System.

expression coding standard FACS, respectively insights into how information processing is done in the visual cortex to name only two examples.

One of the key issues in facial motion analysis is the robustness of the system in order to allow its use in any unconstrained environment, under varying lighting conditions and independent of pose. A facial expression analysis system should fulfill therefore following requirements:

- Robustness towards **different subjects:** ethnicity, gender, facial hair, make-up and age.
- Robustness towards **pose variations:** translation, 3D-rotation, scale changes.
- Robustness towards **illumination changes:** artificial or natural light, lighting direction, specular reflections.

Before going into the details of facial expression recognition, let us have a quick look on a generic analysis system that would allow to meet the above stated requirements and depicted in Figure 4. It features four main modules. A test face is first captured, normalized, analyzed and finally extracted facial expressions are classified. Accurate facial expression analysis can only be done with reliable face preprocessing available. Major problems occur with scale and rotation changes of test faces as well as their reliable detection in cluttered scenes. Lighting changes pose further problems and partly occluded faces may get lost in the face detection process. Several approaches have been taken to encounter these difficulties, among them neural network based techniques showed to produce stable results for frontal faces [88] and [14]. Most facial expression analysis methods will require a face in an up-right position, so there should also be mechanisms for detecting head tilts and rotations in order to be able to take appropriate measures to compensate for these in the normalization stage. For example *Rowley et al.* [89] achieved to control in-plane rotations of faces using a neural network approach. Identity, age, gender and ethnicity may be determined and processed independently from facial expression analysis but could constitute further sources of information when trying to distinguish facial expression changes with more accuracy.

# 3 Facial Expression Measurements

Facial expressions are generated by contractions of facial muscles which results in temporally distorted skin texture, eye lids, eye brows, nose and lips. Distorted skin is revealed by wrinkles and bulges as well. Typically, changes in muscular activity are brief, lasting a few seconds, but rarely more than five seconds or less than 250 ms. We would like to accurately measure facial expressions and therefore

Figure 5: Basic emotion samples. These expressions are feigned, note the exaggerated surprise display.

need a useful terminology for their description. Of importance is the *location* of the facial expressions, their *intensity* as well as their *dynamics*.

## 3.1 Facial Action Codes versus Emotion Categories

There are two different basic approaches on how to measure facial expressions. Either classify them using some kind of a facial action code or map them directly into emotion categories. Currently, most computer-based facial expressions recognition systems attempt to classify expressions into a few broad categories. Some of these facial expressions are universal across human ethnicity and cultures and have been identified by Ekman [117], Figure 5 shows some feigned basic emotion expression samples. For each of these "basic emotions" exists a prototypical, innate and universal expression pattern. The evidence for seven universal facial expressions[7] does not imply that these emotion categories are sufficient to describe all facial expressions [126]:

- If automated facial measurement were to be constructed simply in terms of elementary emotional categories, much important information would be lost: blends of emotions, variations within emotional categories[8], intensity[9], conversational signals, and idiosyncratic facial movements [4].
- In everyday situations, people do not often show exaggerated facial expressions and this means that classifying methods based on "basic emotional states" show difficulties when trying to distinguish facial expressions. When it comes to a single basic emotion, there are also sub-categories, for example we are able to clearly distinguish *false smiles*[10], *conversational smiles, anxious smiles, happy smiles and sadistic smiles.*
- Apart from the "basic emotions" there exists various types of affection that are often only clearly interpretable in the complete context: daydreaming, being distrustful, fiendish, haughty, head in backward position, scolding, sneezing, suspicious, etc.
- Facial expressions don't necessarily occur only in combination with emotions, but also as a result of *cognition activities.* Hence, attempting to represent these kind of facial expressions with a set of basic emotional states doesn't make sense.

These findings suggest that an advanced automatic facial expression analysis system would favorably make use of facial action codes in order to represent the multitude of possible facial expressions prior to interpretation. *Facial expression dictionaries* allow to decouple facial expressions from any kind of mental display mapping. This allows for better examination and control of all facial analysis methods involved in a facial expression analysis system, in particular of the classification stage.

[7] Happiness, sadness, surprise, fear, anger, disgust, contempt. See [119] for a review.
[8] E.g. vengeance versus resentment.
[9] E.g. annoyance versus fury.
[10] Not felt smiles, i.e. displayed only with the mouth.

### 3.2 Facial Expression Dictionaries

The *Facial Action Coding System (FACS)* developed by *Ekman* [122] has been considered as a foundation for modeling facial expressions. The first facial animation model to use the FACS notation was that of *Platt and Badler* [85]. FACS uses so called AUs (action units) that express changes of facial expressions in certain facial regions. There are 46 AUs which account for facial expressions changes, describing facial actions such as eyebrows and lip corners being pulled up. Another 8 AUs were defined to code the head position and 6 AUs describe the eye positions. All expressions can be modeled by single action units or action unit combinations. The precise duration of each action is determined and the intensity of each muscular action and any bilateral asymmetry may be rated. FACS is a purely *appearance-based* description framework that does not carry any information on the mental states associated with the expression. However, experiments have shown that a high correlation exists between the FACS and specific facial emotional patterns [118].

A computer vision system that could detect the activity of specific AUs on a face would be particularly beneficial, since these can be mapped to facial expressions and generate hypotheses on possible emotions. Facial actions often are displayed in combination. With training it is possible though to control each of these actions separately. FACS is currently used as a research tool in several branches of behavioral science. It has been applied successfully in psychotherapy research, for example to evaluate therapies such as treatment programs for stutterer [112]. Another application example is the research done on prototypical affective micro-sequences in psychotherapeutic interactions [111]. FACS is not the first classification approach but the most widely adapted one among psychologists and therefor also the best understood classification scheme.

Izard's *Maximally Discriminative Affect Coding System (MAX)* [127], is an alternative facial action coding system where the units are formulated in terms of appearances that are relevant to eight specific emotions rather than in terms of individual muscles. Unlike FACS, MAX does not exhaustively measure all facial motions. Face actions specified by MAX comprise a subset of facial actions contained in the FACS facial expression dictionary.

*Essa and Pentland* [39] introduced yet another alternative to FACS called FACS+. They criticized that FACS would be too complex and that the dynamics of expressions, rather than the detailed spatial deformations are important in expression recognition. Essa and Pentland also claimed that the temporal information would be completely missing from FACS. FACS+ introduces probabilistic facial muscle activations associated with each expression, using a detailed physical model both of the skin and the muscles. Figure 6 compares an observed facial action versus a FACS coded representation. *Bartlett et al.* [4] argued that this criticism seems to be based on a confusion between FACS itself and computer graphics systems that implement aspects of FACS. One could think that important subtle spatial distortion and motion may get lost in the FACS coding process and therefore a later accurate mapping towards mental states would not be achievable. On the other hand, FACS may allow for better generalization, suppressing inter-personal variations in physiognomy of faces as well as noise.

*Yacoob and Davis* [105] developed a mid-level symbolic representation motivated by psychological considerations. Recognition of six universal facial expressions (happiness, sadness, anger fear, disgust and surprise), as well as eye blinks, were demonstrated on a large set of image sequences. Expressions were further sub-categorized into three temporal parts: beginning, epic and ending. Similarly, *Padgett and Cottrell* [78] used a dictionary based on basic emotional states and pairs of basic emotions in order to interpret facial expressions.

*Black and Yacoob* [17] described facial motion in terms of a small number of parameters such as translation and divergence for each facial feature. These parameters are intuitively related to the motion of facial features during facial expression displays and are estimated for each feature separately. Mid-level predicates can then be derived that characterize the motion of these features.

Finally, the MPEG-4-SNHC (Synthetic Natural Hybrid Coding) is a standard that encompasses analysis, coding and synthesis of faces. Here, shape, texture and expressions of the face are generally controlled by the bitstream containing instances of Facial Definition Parameter (FDP) sets and/or Facial Animation Parameter (FAP) sets. These are closely related to muscle actions and contain 3D

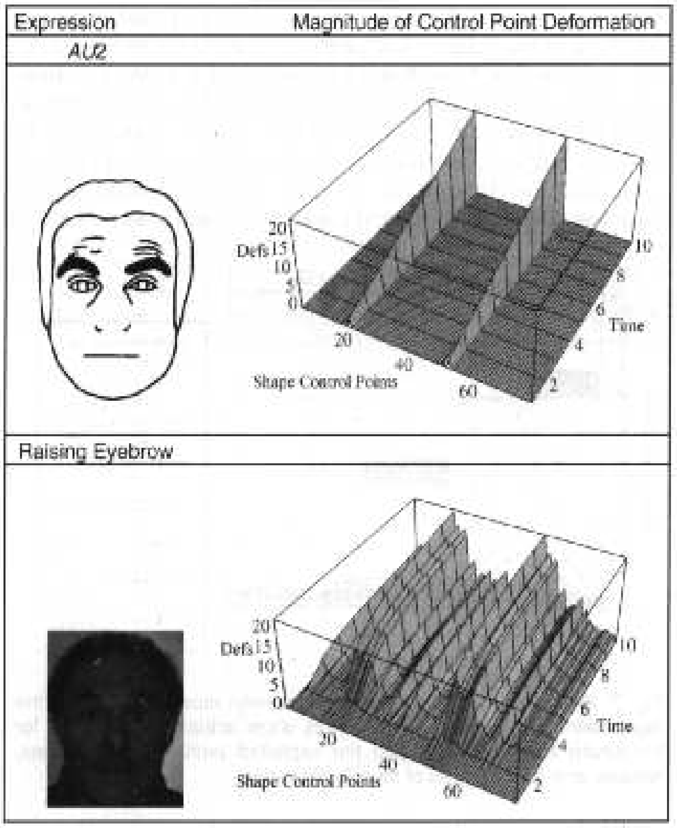


Figure 6: *FACS deformation versus observed deformation for the raising eyebrow expression.* On top is given the surface plot of FACS AU2 with changing intensity and on the bottom an actual video sequence of raising eyebrows [46]. Since facial expressions activities and facial expression intensities are quantized with FACS, the above shown signal is deprived of lower amplitude components. This may lead to unrealistic synthetic re-animation results but is probably sufficient to accurately describe and interpret the majority of all possible facial displays.

feature points[11] and optionally a 3D mesh with texture coordinates if appropriate, texture images and other information such as hair style, glasses, age and gender. Upon construction, the Face object contains a generic face with a neutral expression. This face can already be rendered. It is then immediately capable of receiving the parameters from the bitstream, which will produce animations of the face caused by expressions and speech. If FDPs are available, they can be used to transform the generic face into a particular face determined by its shape and optionally texture. If desired, a complete face model can be downloaded via the FDP set as a scene graph for insertion in the face node, see [58]. *Eisert and Girod* [37][38] used a set of facial parameters based on the MPEG-4 standard and which are determined from two successive video frame applying a hierarchical optical flow based approach.

### 3.3 Facial Expression Intensity

Facial expressions occur with changing intensities. For example the degree of smiling is communicated by the magnitude of cheek and lip corners raising as well as wrinkle display. It is important to be able to approximately estimate different facial expression intensities. Approximately because there are inter-personal differences between the maximum possible intensity amplitudes of facial expressions and it is therefore difficult to determine the exact absolute facial expression intensity of a given person and expression without referring to a neutral face of that person. Nonetheless it is possible for human beings to judge facial expression intensities of unseen persons, probably achieved by using generic neutral face models learned by experience. Of importance is also the *relative facial expression intensity* in a dynamic context, in other words, temporal intensity changes, see next paragraph.

FACS allows AUs to be produced at different levels of intensity[12]. Originally there were *three* intensity levels, where $x$ denoted the minimum perceptible AU intensity, $z$ the maximum possible peculiarity which is dependent on a specific AU. Everything in between is coded with an $y$ intensity. Later, *five* intensity levels were introduced with minimum requirements for coding attributed to each AU: with $a$=*trace*, slightly below the minimum requirements, $b=x$ minimum requirements fulfilled, $c=d=y$ and finally $e=z$ the maximum possible AU intensity. In addition, the lists of AUs with defined intensity coding was enlarged in comparison to the original system proposed by Ekman et al.

Facial expression intensities are difficult to measure because each AU manifests its magnitude of expression in a different way and FACS doesn't specify exactly how it should be measured. Unfortunately a criteria, that would allow for a precise description on how to encode facial expression intensities, is missing. Manual coding of image sequences is time a time consuming task, that's why five intensity levels are usually only used by people who do EMFACS [125] coding. EMFACS is, in contrast to FACS not all encompassing, because only those AUs are coded, that Ekman and Friesen found to be relevant in basic emotions. Furthermore, EMFACS captures not the whole evolution of the intensity dynamics, as one only starts to record shortly before the intensity apex is reached.

### 3.4 Facial Expression Dynamics

Psychologists do believe that a dynamic description of facial expressions is important for analyzing both slow and rapid facial signals [128]. One example for the relevance of temporal aspects is the fact that most people are very sensitive to false timing of facial expressions, when evaluating the truthfulness or deceitfulness of an emotional display, e.g. abrupt endings or beginnings of a smile. Timing for example contains important information for discriminating genuine emotional smiles from false smiles produced voluntarily such as when smiling for the camera [121]. Huang [54] demonstrated that the relative timing of mouth and eye motion in synthesized expressions can shift perception of a smile from a genuine smile to a voluntary smile or a sneer. Nishio [74] studied the temporal difference in eye and mouth movements for the classification of smiles and postulated a hypothesis on how

[11] E.g. mouth corners and contours, eye corners, eyebrow ends, etc.

[12] Another term used in the literature [4] is facial expression *magnitude* which doesn't lead to possible confusion with the image intensity.

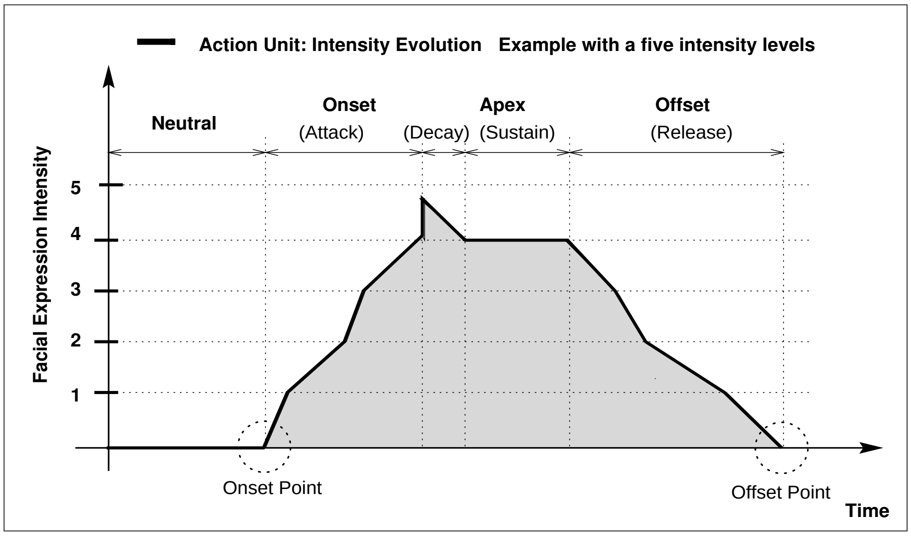


Figure 7: FACS facial expression intensity coding.

we perceive smiles as either pleasant or unpleasant: If the mouth begins to move prior to the eyes, the observers' attention is drawn to the mouth and a smile is perceived as pleasant and vice versa. Although the importance of "correct" timing is widely accepted, only a small number of studies have investigated this aspect systematically and mostly for smiles. Empirical studies on the dynamics of facial expressions of emotions are relatively scarce [129][110], but a few theoretical assumptions have been formulated by emotion theorists. One approach is situated in the tradition of "discrete emotion" theories [120]. Another approach is upheld by "cognitive emotion" theories [134]. In a judgment study [137], FACE was used to test these different theoretical predictions on the dynamics and differentiation of emotional facial expressions by presenting two versions of animated stimuli (simultaneous versus sequential action unit onset).

FACS achieves facial expression coding on a frame by frame basis without the use of inter-frame related information. Insofar it is a *static* coding approach, and does basically spatial compression, subdividing the face into action units that are most relevant in representing the underlying muscle activity. No dynamic information is lost though, as the activation of action units and their intensity respectively are registered in the corresponding succession of frames. Further on, FACS allows to annotate facial actions with attributes describing the evolution over time: onset, start of apex, end of apex and offset[13] . Figure 7 shows a sample intensity evolution of a FACS action unit. FACS encodes facial actions in each frame separately, therefore the *temporal resolution* that can be achieved depends on the frame-rate of the video system applied. Usually, this means 25 frames per second for PAL and 30 frames per second for NTSC. These temporal resolutions have shown to be accurately enough for the detection of rapid facial signals such as eye blinks.

## 3.5 FACS Requirements

Facial expression extraction methods should be carefully chosen according to the kind of facial expression coding scheme that is used in a later classification stage. The best suited approach for facial expressions analysis depends on the purpose of the system, namely if we would like to *animate synthetic faces*, directly *map facial expression to basic emotion categories* or *code facial expressions* with a dictionary such as FACS. This has the advantage of decoupling the facial expressions recognition stage from the interpretation stage. We focus here on the latter approach which requires:

[13] Alternative expressions: *attack* or *initiation* for onset, *sustain* for pex, *relaxation* for offset.

| **Transient Facial Features** | **Intransient Facial Features** |
|---|---|
| • Wrinkles | • Furrows due to aging |
| • Iconic changes of eyes[1] | • Eyes (contour), eyebrows, mouth, nose |
| • Iconic changes of mouth[1] | • Tissue Texture[2] |
| | • Facial hair |

Table 1: *Facial features that change form and location with muscle activation.* [1]Appearance and disappearance of facial features (eye balls, teeth, wrinkles). This lead often to abrupt changes as in the case of eye blinks. [2]Distortion of not trackable regions can be made visible using markers for highlighting.

- **High spatial accuracy:** The mouth is the most discriminating facial feature followed by the eyes and the eyebrows. Wrinkles and furrows are also part of the so called facial features. There are two different types of wrinkles though: *transient wrinkles* that stem from muscle contraction and are of high validity and *intransient wrinkles* or furrows due to old age. Wrinkles are important for FACS coding and it is not sufficient to focus only on the main facial features but holistic analysis should also be integrated. Transient wrinkles appear most often on the forehead, around the mouth and the eyes. Table 1 summarizes facial features that need to be analyzed for deformation, distortion and displacement when doing facial expression recognition.

  *Kaiser and Wehrle* [128] proposed a facial expression analysis system that works with a resolution of $352 \times 288$ or better. Note that their systems is based on the tracking of facial markers that can be tracked rather easily, even on blurred images. On the other hand, systems that operate marker-independently need to extract also subtle facial display changes, e.g. wrinkles, so a minimum resolution should be guaranteed. *Lien et al.* [64] used images of size $490 \times 640$, cropped to $417 \times 385$, whereas *Bartlett et al.* [4] analyzed upper and lower face images at a resolution of $60 \times 90$. Both systems employed FACS coding of facial expressions. A standard CIF resolution of $352 \times 288$ showed to be sufficient for FACS coding done by human experts.

- **High temporal precision** is needed to capture rapid facial signals such as eye blinks and therefore coding has to be done at full video rate with at least 24 frames per second.

- **Face segmentation:** FACS divides faces into an upper and lower part, where the former contains the face regions from the nostrils upwards and the latter from the nostrils downwards. Upper and lower face regions can be treated separately because there are relatively few interactions of facial actions between the respective regions.

As we have highlighted before, facial expression dynamics are important for emotion classification. Even though many approaches presented in the literature employ dynamic analysis methods, it is not mandatory for FACS coding. Dynamics of facial expressions on the other hand are not affected by FACS coding and can be retrieved for later interpretation. Note that methods that directly map facial expressions into basic emotions may be less demanding concerning spatial accuracy and may work reasonably well using only a few facial features to estimate emotion and thus allowing for lower image resolutions. Indeed, *Zhang* [109] suggested that $64 \times 64$ pixels should be sufficient to analyze a face for facial expressions.

## 3.6 Face Characteristics and Measuring Aspects

Age, gender, ethnicity and identity influence facial expression displays. The resulting variations should be taken into account in order to reliably capture facial changes:

- **Age:** old aged peoples faces contain intransient wrinkles which lead to different facial tissue deformation in conjunction with facial expression displays in comparison to what can be observed on smooth skin. Young peoples faces and especially baby faces tend to be smooth. This, in combination with an insufficient image resolution, could degrade tracking results due to missing texture.

- **Gender:** Female faces are generally smoother than male faces which leads to the same problems as in the case of young peoples faces. In addition, male faces tend to be more often covered by facial hair. Hair occludes underlying tissue which may impair facial expression extraction algorithms because tissue distortion is displayed different on facial hair covered areas or not even be visible at all.
- **Ethnicity:** Facial expression extraction methods should be independent of physiognomy.
- **Identity:** Adaptation to intra-personal variations may improve facial expression extraction results.

Facial actions can be described in a purely appearance-based manner using FACS, where muscle activities are represented indirectly. Another possibility is to estimate *facial muscle contractions* from skin and facial feature deformations. The latter approach is well suited for the animation of synthetic faces, the former allows for direct mapping of facial expressions to mental states, because it is the effect of the muscles that are interpreted in non-verbal communication and *not* muscle activities. Of interest is the robustness of a given coding scheme with the presence of skin distortions and partly occluded faces that occur due to out-of-plane rotations. Note that only frontal faces can be coded with FACS. Absent or distorted AUs may lead to activity and intensity estimation errors.

# 4 Facial Expression Analysis

Most research activities in the field of automated facial expression analysis and recognition has taken place in the last ten years, starting in the beginning of the nineties. Table 2 lists facial expression analysis systems referenced in this paper.

Various methods and techniques have been proposed to encounter the complex task of facial expression analysis. Figure 8 gives an overview of the various facial expression analysis stages found in a complete facial expression analysis system. First, a face, once it has been located in a given scene, needs to be normalized to encounter environmental changes. Then it is segmented according to the facial features of interest. Next follows the extraction of facial expressions as well as a suitable representation, using a certain number of parameters. Finally, facial expressions are classified with the aid of a facial action coding dictionary. Furthermore, facial expressions may be interpreted, where a mapping of coded facial expressions towards mental states takes place. In this chapter, we don't discuss facial expression interpretation though, as this survey mainly focuses on facial expression analysis and recognition. Following paragraphs are presented according to the outlined processing stages, introduce some of the methods most often used and highlight their characteristics.

## 4.1 Face Normalization

Prior to face segmentation and facial expression analysis, incoming test faces are normalized. This necessary in order to ensure that environmental dependent variables such as head motion, face size and orientation as well as lighting conditions do not interfere with facial expression recognition. Following normalization procedures need to be addressed:

- **Pose normalization:** Facial expressions analysis is rather sensitive to pose variations caused by in-plane rotations and especially out-of plane rotations. These lead to false distortions and partly occluded faces. Most systems presented in the literature are only able to analyze frontal faces that are shown in an up-right position and pose variation changes are controlled by manually *aligning* test faces. Fully automated in-plane position normalization of faces was demonstrated

| Motion-based Approaches | Distortion-based Approaches | Hybrid Approaches |
|---|---|---|
| Terzopoulos & Waters[97] (1990) | Suwa et al. [95] (1978) | Lien [62] (1997) |
| Mase [71] (1991) | Kaiser & Wehrle [128] (1992) | Cohn et al. [25] (1998) |
| Cottrell & Metcalfe [30] (1991) | Lanitis et al. [59] (1997) | Bartlett [4] (1998) |
| Moses [73] (1994) | Padgett & Cottrell [78] (1997) | |
| Rosenblum et al. [87] (1994) | Pighin [84] (1998) | |
| Yacoob & Davis [105] (1996) | Lisetti [65] (1998) | |
| Essa & Pentland [39] (1994) | | |
| Kobayashi et al. [57] (1995) | | |
| Black & Yacoob [17] (1995) | | |
| Darrell et al. [33] (1995) | | |
| Bascle & Blake [7] (1997) | | |
| Oliver et al. [75] (1997) | | |
| Eisert & Girond [38] (1997) | | |
| Chen et al. [23] (1998) | | |
| Wang et al. [101] (1998) | | |
| Zhang [109] (1998) | | |
| Lyons [70] (1998) | | |
| Hua [53] (1998) | | |
| Otsuka & Ohya [77] (1998) | | |
| Hong et al. [51] (1998) | | |

Table 2: *Selected facial expression analysis systems* categorized according to the facial expression extraction approaches employed. The years given in parentheses don't necessarily indicate the first available publication of a certain author on facial expression analysis.

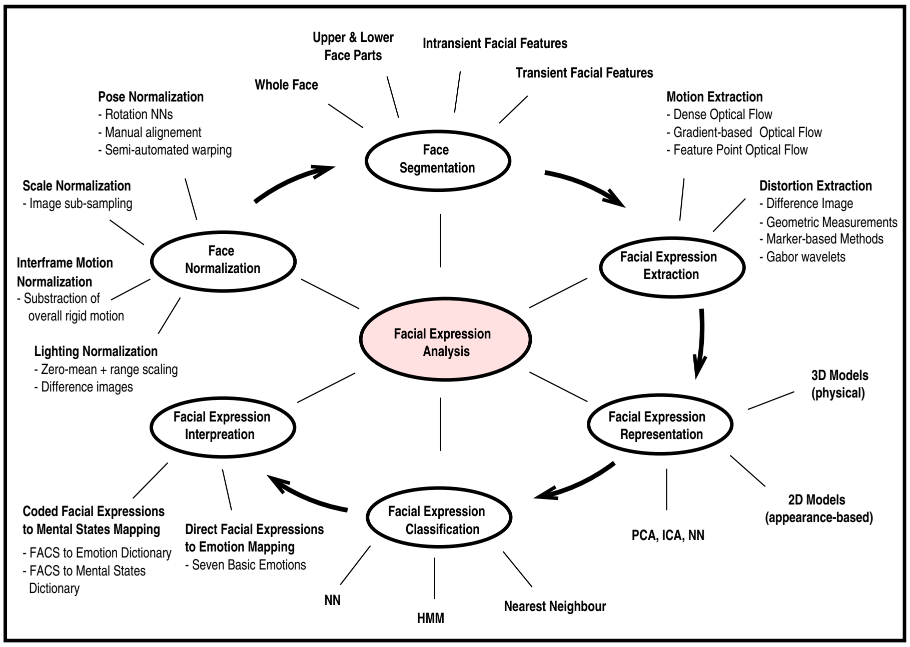


Figure 8: Overview of facial expression analysis methods and their interaction in a complete facial expression analysis system.

in [89]. Faces were rotated from an arbitrary into an upright position. An open problem is the alignment of out-of-plane rotated faces, as parts of the faces can be occluded or distorted and may only be recovered insufficiently in a subsequent frontal face type facial expressions analysis stage. Out-of-plane rotation normalization could be handled though by methods that were proposed in [15][100]. The choice of how to treat pose variations is of great importance, because it directly influences the facial extraction methods that can be applied.

- **Lighting normalization:** Lighting variations can be addressed in a preprocessing stage, independently of facial expression analysis. There are some methods that are more sensitive than others to lighting variations, e.g. optical flow. One frequently used approach to overcome changes in amplitude is to first compute the zero-mean of a given test image and then scale it into a given range that is independent of lighting scale changes in the original test image. Note that lighting changes are not much of an issue when using difference-images [6]. There, illumination changes are automatically eliminated by subtracting a reference frame from a test frame, as both frames are recorded under the same environmental conditions. For more details on difference-images see further below. Difficult to solve is the problem of partly lighten faces, where half of the faces are invisible. Some kind of face models would be needed to recover the missing parts of the face. Currently, this issue hasn't been addressed in facial expression analysis systems found in the literature.

- **Scale normalization:** Scale variations can also be addressed independently of the facial expression analysis stage, but may be integrated into a single framework together with the recovering of pose variations. Facial expressions could then be analyzed at different scales. Due to changes in resolution, facial expressions can be less reliably distinguished on faces that are further away. Hence a scale-specific treatment may yield higher accuracy over a multitude of scales. Multiple scales analysis can be achieved by detecting faces on subsampled versions of a given test image [88][14]. The scale factor of detected faces is then known and faces may be warped to a constant size for facial expression analysis.

- **Interframe motion normalization:** Rigid and non-rigid facial motion separation is especially important in the case of motion analysis of test sequences. *Lien* [64] applied an affine normalization transformation that included translation in addition to scaling and rotation factors. Motion separation was achieved by subtracting the total facial motion from the motion of selected facial feature points, where the latter carry only rigid motion components.

Note that in most facial expression analysis systems pose, interframe motion and scale normalization is achieved by affine transformations of test faces based on manually entered positions of distinctive facial feature points such as the eye locations (either center or medial canthus) and/or the lower ending of the nose (philtrum), e.g. see [64] and [4].

## 4.2 Face Segmentation

Face segmentation takes place prior to facial expression extraction. It allows to focus on facial areas that are good indicators for the distinction of facial expressions. A face can be subdivided according to following areas of interest:

- **Intransient facial features** such as eyes, nose, eyebrows and the mouth are probably most important for the display of facial expressions. Furrows due to old age are another type of intransient facial features. Typical representatives of locally acting facial feature processing approaches can be found in [17][78].

- **Transient facial features** such as wrinkles occur often with facial expressions displayed with high intensities and are mostly appearing on the front and the regions surrounding the mouth and the eyes. These areas may be processed separately, where wrinkles are examined locally

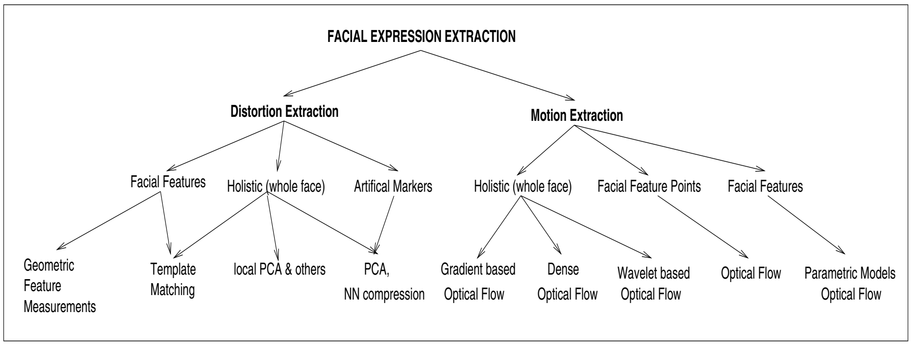


Figure 9: Overview of distortion and motion based extraction methods.

with the aid of image intensity profiles [4] or holistically, applying high gradient components analysis as presented in [64].

- **Whole face**: A given test face is processed holistically, in other words, the whole face is taken into account. In this category fall difference-image [4], dense optical flow [44][64] and appearance-based [59] approaches. Furthermore, a strategy often applied in conjunction with FACS coding is the division of test faces in upper and lower face part as has been described above.

Holistic gray-level information appears to play and important role in human face processing [139][141]. This suggests that facial expression analysis should be done holistically or in combination with locally acting methods. This assumption is supported by the results obtained with automated face recognition [21][31]. Holistic methods are good at representing prevalent facial expressions, whereas local methods applied on facial features detect subtle changes in small areas.

## 4.3 Facial Expression Extraction

Facial expression extraction methods can be categorized according to what they extract from test faces. We distinguish approaches that extract interframe evolution or *motion* and others that focus on *spatial distortion* of facial tissue and features. Distortion of facial tissue leads to transient facial features, iconic changes and deformation of intransient facial features. Facial feature processing takes place either in a *holistic* way where the whole face is processed at once or *locally*, focusing on *intransient facial features*. Figure 9 gives an overview of methods that were applied for distortion and motion extraction.

Colors allow for better tracking of intransient facial features, e.g. red lips. Non perceptible facial changes can be made visible by either painting or affixing marker points on facial regions that show no apparent texture, but directly represent underlying muscle contractions. This is of special interest in the medical domain, where accurate measurements of muscle activities and contractions is of importance. Another possibility is to highlight natural facial features with color in order to allow for a better tracking. This can be achieved by applying distinct colors not present in the face, e.g. using green or blue color.

Other approaches try to extract image features according to spatial frequency. High frequency parts, occurring in facial areas with e.g. wrinkles, often reveal facial expressions. Image analysis in the spatial frequency domain can be achieved with high-pass filtering, gradient-based filter methods or by convolution with Gabor wavelets. The latter method is of special interest because it detects edges over different resolutions. It has been shown, that the human brain detects edges in a similar way.

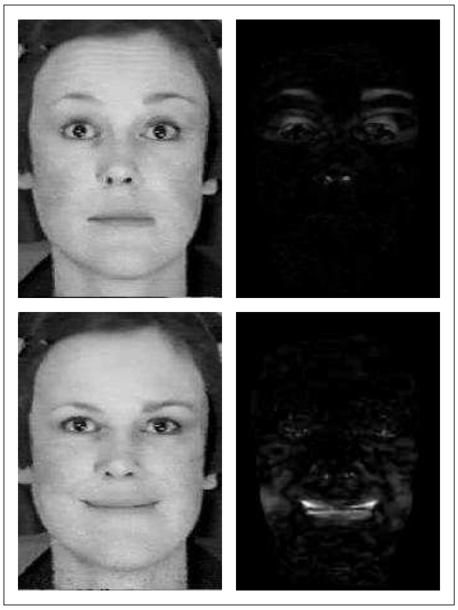

Figure 10: On the left hand side are shown samples of incoming test faces, on the right hand side *difference images* obtained by subtraction of the test faces from a neutral face of the same subject.

#### 4.3.1 Distortion Extraction

Human experts are able to FACS code still images, but a *neutral face* for comparison is necessary, as some AU combinations can only be coded reliably if interframe motion is available. This suggest that computer implementations may also make use of reference frames containing neutral faces. Distortion of facial tissue and facial features can be made visible by comparing a given face with a previously registered neutral face. This can be achieved straight forward by using *difference images* that are created by subtracting a given face frame from a frame containing a neutral face of the same person. Figure 10 shows some sample difference images and the corresponding test frames. Difference images have the advantage of focusing on facial expressions, ignoring other face characteristics that may hinder facial expressions analysis. They reduce also image variability and significantly diminish surface-level differences in appearance. *Bartlett* [6][4][5] successfully applied holistic whole-face difference images with various facial expression *representation* methods that are introduced in the next paragraph. Note that the spatial pattern of facial distortions can also be extracted from a given test face without the aid of difference images. This has been demonstrated in [78] using the PCA representation directly applied on face images.

Another approach to distortion analysis are spatial frequency filtering methods. *Gabor wavelet* filters closely model the receptive field properties of cells in the primary visual cortex [147][34]. Endpoints of lines as well as edge borders can be detected over several scales and in different orientations. These reveal much about facial expressions as both transient and intransient facial features mostly give raise to a contrast change with the ambient facial tissue. Such filters remove also most of the variability in images that occur due to variation in lighting and contrast. Gabor wavelets filter approaches showed to perform well with facial expression analysis [109][4][70][51]. *Bartlett* [4] mentioned that she obtained two different sources of evidence that high spatial frequencies are important for classifying facial actions. Spatial smoothing of optical flow degraded performance by more than 30%. Furthermore, classification with only the high frequency parts of the Gabor representation was superior to classification using only spatial frequencies. A similar result was obtained for PCA jets. Figure 11 displays the results obtained from the convolutions of Gabor wavelet kernels with two face images at different resolutions. *Moses* [73] did facial expression analysis using only the mouth as

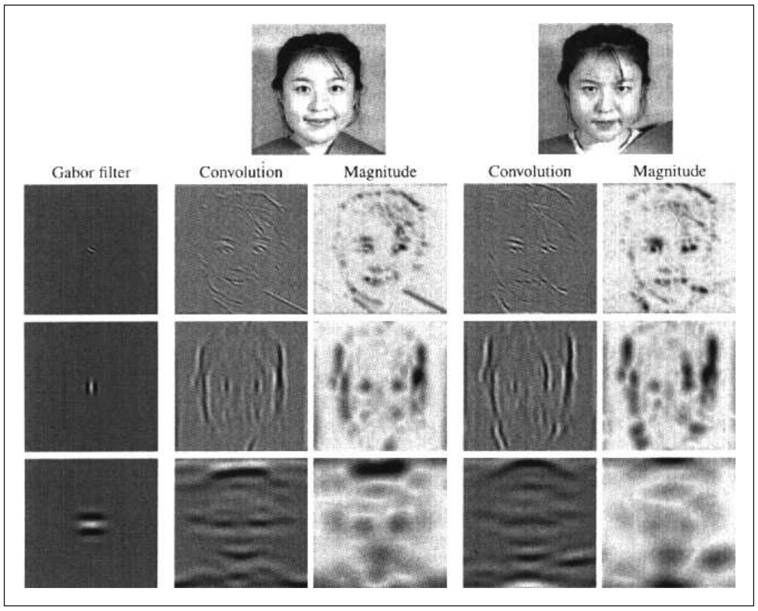


Figure 11: *Gabor Wavelet Representations* [109]. Depicted in the leftmost column are three Gabor kernels with distinct orientations and resolutions. These are convoluted with the test images shown on the upper row.

input, focusing on the valley between the lips as a discriminant feature which is extracted by gradient filtering. Although the mouth is an important facial feature for expression analysis, it is limited and cannot be used to distinguish all possible facial expressions.

Yet another type of spatial pattern analysis approaches are *marker-based* methods. Facial features constitute natural markers and reveal underlying muscle activities through deformation. They can be highlighted to increase contrast and allow for more reliable tracking results over a sequence of frames as has been demonstrated in [7][8]. Previously discussed extraction methods rely on perceptible changes in the face. It is possible though to represent muscle activity in a more direct and more reliable way, namely by measuring distortion in areas where underlying muscles interact. Unfortunately, these are mostly skin regions with relatively poor textures. Highlighting is necessary and can be done by either applying color to salient facial features and skin or by affixing colored plastic dots that serve as markers.

*Suwa et al.* [95] did a preliminary investigation on automatically analyzing facial expressions from a series of movie frames. They approached the problem from a muscular point of view. Twenty highlighted spots in the face were identified and their movements tracked in a sequence of video frames. They were then compared with the movements from prototype patterns for different expressions to determine the expression of the face. Currently, marker pattern analysis approaches deliver the most accurate results of all facial expression analysis methods and FACS coding with a performance comparable to human experts was achieved by *Kaiser and Wehrle* [128], see also further below. Figure 12 shows a comparison of two marker based approaches, namely using either affixed markers, or natural markers called feature points which are regions on the face, that are meaningful and can easily be tracked by feature point optical flow methods.
Table 3 finally compares some of the previously discussed facial expression extraction methods.

### 4.3.2 Motion Extraction

In contrast to distortion-based approaches, motion-based facial expression extraction methods analyze inter-frame evolution and measure the displacement of facial regions and facial features over several

| Method | Characteristics | Assumptions | Advantage | Disadvantage |
|---|---|---|---|---|
| **Difference Images** | Subtraction of face image to be analyzed by neutral face of the same subject. | Neutral face images of a given subject available. Face position, orientation and scale known with high precision. | Focusing on distortions caused by facial expressions. Reduced influence of lighting changes and individual differences. | Availability of neutral face. Precise alignment of neutral face with face to be analyzed necessary. |
| **Geometric Features** | Geometric measurement of facial features. | Facial features with relatively high contrast present. | Relatively simple measurement techniques. Methods that complement other facial expression extraction methods, especially for the measurement of wrinkles. | General performance well below other methods. |
| **Marker Tracking** | Tracking of marker patterns on frontal faces. | Marks represent accurately underlying muscle activities or facial distortions caused by facial expressions. Markers should be visible all the time (frontal face images!) and reliable distinction from the face always possible. | System available with FACS coding performance comparable to human experts. | Markers need to be affixed or painted on the face. |
| **Gabor Wavelets** | Gabor wavelets can reveal high spatial frequencies in faces caused by edges and lines. | Gabor wavelet kernels of adequate resolution with regard to the current scale of the face available. A few discrete angles and resolution are sufficient to describe all facial features. | High frequency components carry significant information about facial expressions. Robustness towards lighting changes. | Additional computational overhead. |

Table 3: Characteristics of Difference Image, Frequency and Marker based Facial Expression Extraction Methods.

Figure 12: Left-hand side: *Marker-based approach.* Marker point locations used by *Kaiser and Wehrle* [128], highlighted with affixed dots. Right-hand side: *Feature Point Tracking:* Locations of selected facial feature points, marked with crosses. They were chosen to reflect best the motion of natural facial features. Samples taken from *Lien* [64].

frames in a temporal context. Most recent research has focused on *optical flow* based approaches for the discrimination of facial expression displays. These approaches are based on the assumption that muscle contraction causes deformation of overlying skin. Specifically the velocity and direction of pixel movement across the entire face or within windows selected to cover certain facial regions are computed between successive frames or between a reference frame containing a neutral face and given test frame. *Mase* [71] did pioneering work applying optical flow estimating the activity in 12 of the 44 facial muscles. For each muscle he defined a window in the face image and an axis along which each muscle expands and contracts. The mean similarity of the flow vectors inside the window to this axis provided a coarse estimate of the activity of the muscle. Dense optical flow motion was quantified into eight directions. The resulting motion vectors were computed in six predetermined regions (eye, nose, mouth and eyebrows) which were then categorized into a simplified version of the FACS rules. Optical flow techniques can be categorized as follows:

- **Dense optical flow** tracking approaches are often employed for facial expression analysis [108][64][39][53][37][4]. These approaches allow for a dense estimation of the motion at different points in the scene. The motion vector field is computed using a temporal and a spatial gradient. The main drawbacks of these approaches are their lack of robustness with changing lighting conditions as well as iconic changes. Furthermore they demand prodigious computational requirements. These methods makes the assumption of constant radiance of the object and the movements are constrained to be small. These assumptions are often violated with faces in real-world scenes, where varying non-uniform lighting conditions are common and eye blinks lead to iconic changes.

- **Facial feature point** approaches were used in [78][79][101][84]. For best results (no tracking loss) these points should be selected around intransient facial features. Difficult is the initial placement of the feature points, which is often done manually. Feature points are used in locally acting approaches, as they allow to track small facial regions. On the other hand, feature points cannot reliably track skin motion in areas with low contrast and are therefore only applied near intransient facial features, where a minimum of texture is guaranteed, see also Figure 12.

- **Gradient-based optical flow** was used in [4] and gives reliable estimates only at points of high gradient such as moving edges. This method is computationally efficient since it is based on the local estimation of image gradients that can be computed directly from a pair of images by applying for example horizontal and vertical Sobel edge filters for the spatial gradients, respectively taking the difference of image-pair pixel intensities for the temporal gradients.

Different *optical flow types* were used for the purpose of facial expression analysis. *Lien et al.* [63] employed *Wu's* [103] approach of dense optical flow to estimate the entire facial motion using scaling functions and wavelets from *Cai and Wang* [22] at multi-resolution levels to provide window functions of multiple support sizes, capturing both local and global facial characteristics. Image motion in large regions is estimated with large windows, the opposite is true for small regions. This approach is different from the traditional coarse-to-fine pyramid methods and takes into account detailed information at various image resolutions which leads to more accurate motion estimation. *Bartlett* [4] used *Horn's* [52] and *Singh's* [93] optical flow, *Essa et al.* [44] *Simoncelli's* [92] optical flow*, Yacoob et al.* [108][107][106][105] and *Rosenblum* [87] *Abdek-Mottaleb's* [1] optical flow.

The question which optical flow method delivers best results certainly depends on the given task at hand. *Barron et al.* [2] compared several classic optical flow methods on both real and synthetic image sequences. They found the local differential first-order method of *Lucas and Kanade* [66] as well as the local phase-based method of *Fleet and Jepson* [49] to be most reliable. Further on, Barron et al. highlighted the importance of temporal smoothing as it reduces noise. Local methods showed to be more robust, especially with gradient measurements errors caused by quantization noise.

### 4.3.3 Distortion versus Motion Extraction

Distortion-based extraction methods try to locate spatial patterns in static images in order to estimate facial expression activities. On the other hand, motion-based extraction algorithms measure interframe changes of spatial patterns. Difference image based approaches belong to the former category, although a reference frame is employed to extract facial changes. Motion is mostly computed in a temporal context between several frames. This allows to represent the evolution of facial expressions. Note that motion can also be computed between a neutral frame and a given test frame, similarly to what is done in the case of difference image based facial expressions analysis, see [4]. This would also allow to analyze still image with motion-based approaches, where usually a sequence of test images is needed. There is a difference between these two approaches though, as in motion-based methods, amplitude and direction of pixel elements are determined, whereas with difference images only pixel-wise amplitude differences are at disposition. Furthermore, the latter amplitude is not constant, as it is made up of the difference of pixel intensities at a given location. A given facial feature pattern may thus appear with changing amplitude when moving across a given subjects face with varying facial expression intensity. In addition it may be dependent on the physiognomy of a given subject. Especially facial hair could have an important influence and degrade the force of expression of difference images.

Note that purely spatial methods focusing on facial distortions only without using neutral faces for comparison may have difficulties with intransient wrinkles, as these could be erroneously interpreted as transient wrinkles. Motion-based methods on the other hand face the problem of discontinuity of facial expressions, encountered with the appearance of wrinkles and other iconic changes. Further on, optical flow-based methods are prone to noise. Methods such as quantizing and smoothing are needed to control the flow and allow for the estimation of a principal direction in a given facial region. Table 4 highlights some characteristics, advantages and disadvantages of optical flow based extraction methods that were applied for facial expression analysis.

At this point the question arises, whether motion or distortion is more important for the distinction of facial expressions. Several authors claimed that motion is more important [44]. Motion extraction approaches may facilitate the discovery of facial expressions, focusing on what is important and also allow to determine the evolution of facial expressions. In addition, it may be important to distinguish feigned from true expressions. However, *Bartlett* [4] pointed out, that humans are not so good at classifying facial expressions from motion and results obtained were only slightly above chance. She obtained a better recognition performance with purely static approaches that were based

Table 4: Characteristics of Optical Flow based Facial Expression Extraction Methods.

| Method | Characteristics | Assumptions | Advantage | Disadvantage |
|---|---|---|---|---|
| **High Gradient Optical Flow** | Gives reliable results only at points where gradient is high. | No overall gain or loss of brightness in the image over time. Overall rigid motion of head small between any two consecutive frames. | Achieves efficiency by greatly reducing image data prior to motion analysis. | Transient features could be missed due to low contrast. Problems caused by iconic changes and specular reflection. |
| **Dense Optical Flow** | Insensitive to subtle ($<$ 2 pixels) | Same as above. | Dense array of motion | Local estimates of |
| **Feature Point Tracking** | Sensitive to both subtle motion ($<$ 2 pixels) and large motion ($>$ 15 pixels) [63]. | Same as above. | Analysis is directed to the objects of interest in the scene; motion is not computed for extraneous background patterns. | Enormous number of relatively complex patterns may be needed to be searched. Velocities can only be reliably extracted at points of large gradient. |

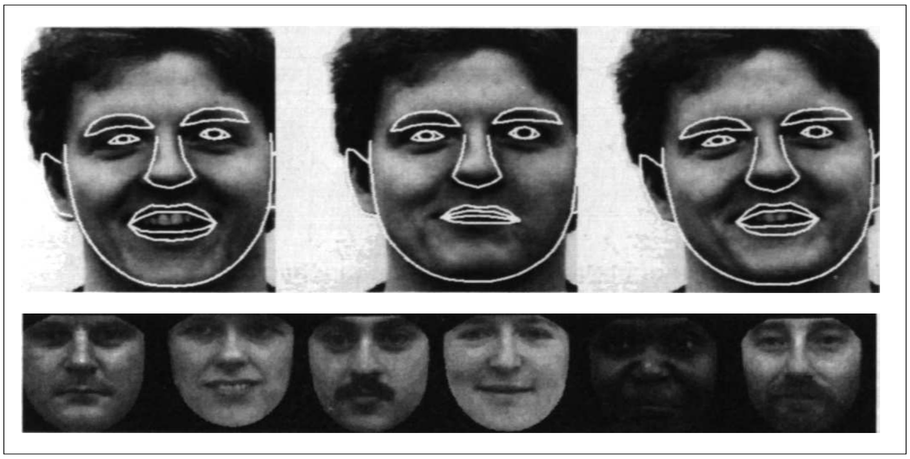

Figure 13: On the upper row is shown a face shape model, adapted to varying facial displays. On the bottom row are given texture models. Source: *Lanitis et al.* [59].

on difference images combined with ICA representations. See also Table 8 for a facial expression recognition performance comparison of both motion and spatial pattern analysis based algorithms.

## 4.4 Facial Expression Representation

Extracted facial expressions need to be represented and described prior to classification. This can be done by using 2D or 3D face models or component analysis approaches. Representing facial expressions with a consistent set of model parameters or as a combination of principal components has the advantage of greatly reducing the data that needs to be classified and the achieved compression allows to focus on important features while reducing disturbing effects caused by noise.

### 4.4.1 Model-based Approaches

Modeling can lead to more stable results, as missing information can be complemented by *a priori* knowledge. We distinguish 2D and 3D models of faces:

- **2D Face Models:** Most model-based systems developed so far have relied on heuristics about shape deformation, which is either too limited or too flexible. This leads to deficiencies in mapping the model to a given image. One of the issues to be addressed in a model-based approach is to choose an appropriate description of the facial structure. Specifically, for visual analysis, one can attempt to estimate muscle actions from the image as done in [71][41]. However, the musculature of the face is complex, 3D information is not readily present, muscle motion is not directly observable. For example, there are at least thirteen groups of muscles involved in the lip movements alone [50].
  *Appearance-based models* can represent faces by using 2D-projections of 3D-objects and do not require to recover 3D information. We distinguish approaches that combine both shape and texture [36][59] as well as approaches relying on whole-face 2D feature point templates that are matched to corresponding facial features [98]. Characteristic of the former method is its dual approach, analyzing faces separately in shape and texture. *Edwards et al.* [36] interpreted face images using whole-face active shape models, based on active appearance models [29]. Figure 13 shows face shape and texture models applied for facial expression analysis and other face processing tasks. Appearance-based model approaches are able to implicitly account for 3D information based on the appearance of the object and allow to separate fairly well different information sources such as illumination, pose and facial expression by supervised learning of

these factors. Appearance changes can be caused by many different factors, including deformation, illumination changes, specular reflections and iconic changes.
*Mixture appearance models* [16] constitute an alternative that may account for these factors simultaneously, breaking the appearance into separate pieces and therefore allow to recover the individual components. The approach is based on the use of a probabilistic mixture models to explicitly represent multiple motions within a patch where a simple extension of the EM-algorithm is used to compute a maximum likelihood estimate for the various motion parameters. *Oliver* [75] applied second-order mixture models to describe both chromatic and spatial distributions of the mouth. In addition image patches around the mouth region were used to obtain better performance, represented with a Gaussian mixture model.
*Parametric motion models* are one way to generate compact representations of intransient facial feature deformations by focusing on the motion associated with the edges of the mouth, nose, eyelids and eyebrows. These are optical flow based methods computed locally in windows centered around intransient facial features and were applied by *Black and Yacoob* [17], *Rosenblum et al.* [87][86] and *Otsuka and Ohya* [77].

- **3D Face Models:** *Samal et al.* [90] reported that *Terzopoulos and Waters* [97] have used a simple form of the models proposed by *Ekman and Friesen* in order to derive a 3D computer model for displaying and analyzing human facial expressions from a video sequence. The physically-based synthetic tissue model was motivated by tissue biomechanics . *Essa and Pentland* [39][45][41][46][44] employed two methods for facial expressions recognition. The first one estimated muscle activities, the other generated whole-face spatio-temporal motion-energy fields. Both methods performed equally. Muscle activity and motion energy distributions were computed with the aid of an dense optical flow framework, using a rather sophisticated 3D face model in combination with Kalman filtering for increased stability.

Common for both 2D and 3D models is a tradeoff between the accuracy of a representation and the number of parameters used in the models. A problem encountered with 3D models is that these require complex mapping procedures that generate heavy computational requirements. In addition, physical models often need to be aligned manually onto a given test face. This is not suitable for an automated facial expression recognition system. Image-based methods can probably be more easily adapted to variations in facial structure and skin elasticity found in the subject population. Table 5 summarizes some of the characteristics of model-based approaches.

### 4.4.2 Component Analysis Approaches

Component analysis approaches are general tools for describing, storing, compressing and analyzing statistical data. They constitute a family of unsupervised techniques which try to represent data in terms of statistically independent variables:

- *Principal component analysis (PCA)* has received much attention in numerous fields. The first application to human faces was pioneered by *Sirovich and Kirby* [94]. Principal component analysis has been successfully applied for holistic facial expression representation based on difference images [4] and is a way of encoding second-order dependencies in the data by rotating the axes to correspond to directions of maximum covariance. This results in a dimensionality reduction. Figure 14 depicts sample PCA components obtained from difference images.

- A related method is *Independent Component Analysis (ICA)* [27][55] is a generalization of principal component analysis. It separates high-order dependencies in the input, in addition to the second-order dependencies. *Bell and Sejnowski* [12] showed that pixel-wise covariances in a image set correspond to the amplitude spectrum, but not the phase spectrum, whereas higher-order statistics contain the phase information. PCA representations focus therefore only on the amplitude spectrum representations. ICA on the other hand includes both the amplitude and

| Method | Characteristics | Assumptions | Advantage | Disadvantage |
|---|---|---|---|---|
| **Local Parameterized Models** | Facial features points are tracked in local windows of size $n$ $x$ m. | No overall gain or loss of brightness in the image over time. Overall rigid motion of the head is small. Spatially bounded local non-rigid motion between any two consecutive frames. | Near frontal view out-of-plane rotated faces can be processed. | Local models are restricted to intransient facial features and it is difficult to cover transient facial features. Iconic changes and specular reflections cause problems. |
| **Appearance based Models** | Faces represented by shape and texture models. | Availability of accurate model based on shape and texture. For speed and reliability reasons: Initial search position is located close to the object to be matched. | Identity, gender and facial expressions recognition possible in one single framework. | Large number of hand-labeled models necessary for FACS coding (AU combinations!). Possible distortion of extracted shape-free models due to non-frontal faces. |
| **Physical Models combined with Optical Flow** | Possible to control optical flow (stability with noisy data). | Availability of accurate physical models. | 3D models allow to stabilize optical flow based approaches, as facial expressions are constrained to the underlying muscle framework (deformation areas, max possible muscle contractions and directions of contraction). | Difficult to construct precise face and head models. Complex transformations from 2D appearance to 3D model data and backwards. Difficult to model transient facial features. |

Table 5: Characteristics of Selected Model based Facial Expression Extraction Methods.

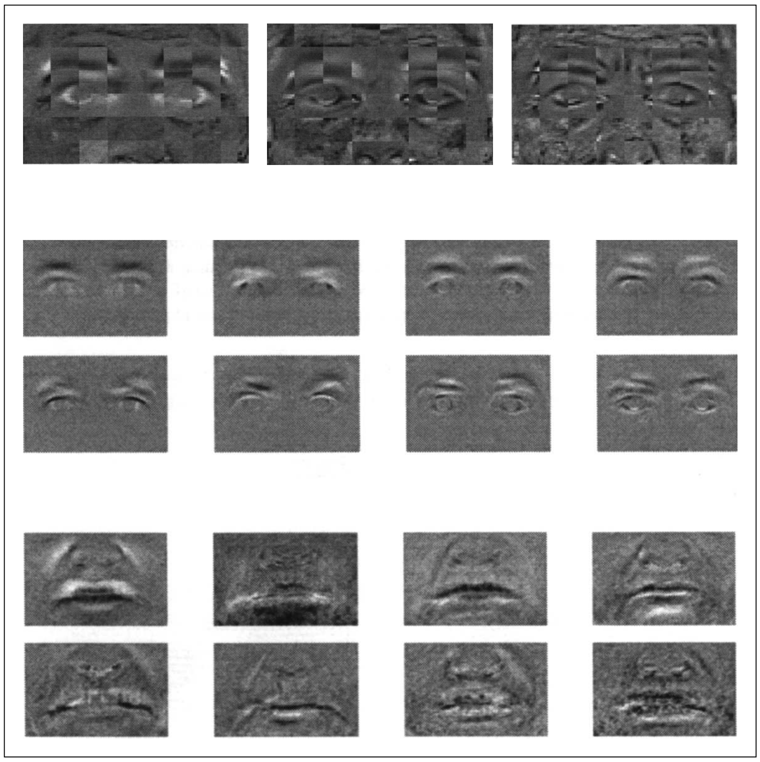

Figure 14: *PCA face representations by Bartlett et al.* [4]. Top row: local principal components of $15 \times 15$ patches in fixed locations on upper face difference images. From left to right: Principal components 1-3. Lower part: First 8 principal components of the difference images from the upper face actions shown on the middle rows, and lower face actions on the bottom rows. Components are ordered from left to right and from top to bottom.

phase spectrum. It has drawn much attention lately. One reason for this is that new efficient learning algorithms have been introduced [11][56].

- *Local Feature Analysis (LFA)* [82] has all the desirable properties of PCA and at the same time explicitly reveals some symmetries not captured by PCA. The method defines a set of topographic local kernels, that are optimally matched to the second-order statistics of the input ensemble. Kernels are derived from principal component axes and allow "sharing" the PCA coefficients to equalize their variance, followed by a rotation to pixel space. *Bartlett* [4] obtained best recognition results with the LFA approach on training sets, but the method did non generalize well, probably due to a too great dimensionality reduction.
- *Fisher's Linear Discriminant (FLD)* [48][10] projects the images into a subspace in which the classes are maximally separated. Between-class scatter is maximized and within-class scatter of the projected data is minimized. FLD assumes linear separability of the classes.
- *Neural networks (NN)* achieve class separation similarly to PCA and can also be used for the neural computation of ICA [56] method.

All these approaches can be employed both holistically where the whole face is processed or locally where mosaic-like patches of the face are taken as input. *Bartlett* [4] compared all of the above described component analysis approaches, see Table 8. Further on, Table 6 presents some characteristics of the most frequently used representation methods for facial expression analysis.

## 4.5 Facial Expression Classification

Facial expressions can be classified either in a *spatio-temporal* and *spatial-only* way. Spatial approaches are well suited when used in conjunction with static facial expression coding schemes such as FACS, whereas spatio-temporal approaches may be more appropriate with coding dictionaries that represent also dynamic information. An example for the latter type of facial expression dictionaries is FACS+ [105]. Table 7 lists some classifiers that were used for facial expression analysis.

### 4.5.1 Spatial Classification

It has been found, that simple approaches such as template matching and nearest neighbor classifiers performed well for the coding of spatial face patterns. Whole-face templates were used to classify dense optical flow [45][4] and feature graphs [101]. More complicated statistical classifiers such as feed-forward neural networks and radial basis function networks were also successfully used to classify facial expressions [78][4][109][65]. Unfortunately, it is difficult to train neural networks with methods such as back-propagation. This is mainly due to the sheer number of possible AU combinations. An alternative constitute compiled neural networks, as proposed in [128].

### 4.5.2 Spatio-Temporal Classification

Hidden Markov Models (HMM), were used to classify facial motion flow in frame sequences combined with [25] and without [77][75] the FACS coding scheme. *Rosenblum et al.* [87][86] computed optical flow on face regions, then applied radial basis function networks to classify facial expressions. A specific neural network was employed for each basic emotion. Radial bases functions allow for the direct representation of prototypical situations of a given application in receptive field centers. The maximum response of a receptive field occurs, when an input image is situated at the same location as the center of the receptive filed. Thus the response degrades in a Gaussian fashion, as the Euclidean distance of the input image to the receptive field center increases. *Rosenblum et al.* extended their RBF neural networks with "context units". These are input units that use self-feedback in order to integrate temporal information from previous frames, when classifying a given frame.

| Method | Characteristics | Assumptions | Advantage | Disadvantage |
|---|---|---|---|---|
| **PCA** | Universal method | Precise face alignment | Representation learned from database without manual intervention. Suppression of noise in test data possible. | It is reported that eigenfaces are highly sensitive to minor variations in the image alignment for the face recognition task [83]. |
| **ICA** | Same as above. | Same as above. | Same as above, plus representation of higher order dependencies such as the relationship among several pixels found e.g. on edges (phase alignment at multiple spatial scales) possible. | Same as above. |
| **NN** | Same as above. | Adequate preprocessing stage and training database available. | Same as above. Possible to represent also non-linear dependencies. | Adequate training examples necessary. |

Table 6: Selected Facial Expression Representation Methods.

| Type | Classification | Expression Dictionary |
|---|---|---|
| **Spatial** | • NN[1] | • FACS |
| | • Template matching | • MAX |
| | • Nearest neighbor | • Parametric models [17][78] |
| **Spatio-Temporal** | • HMM[2] | • FACS+ [105] |
| | | • MPEG-4-SNHC [38] |

Table 7: *Facial Expression Classification Methods applied in previous works.* [1]Neural Network: back-propagation neural networks, Radial basis function neural networks, complied neural networks [128]. [2]Hidden Markov Model.

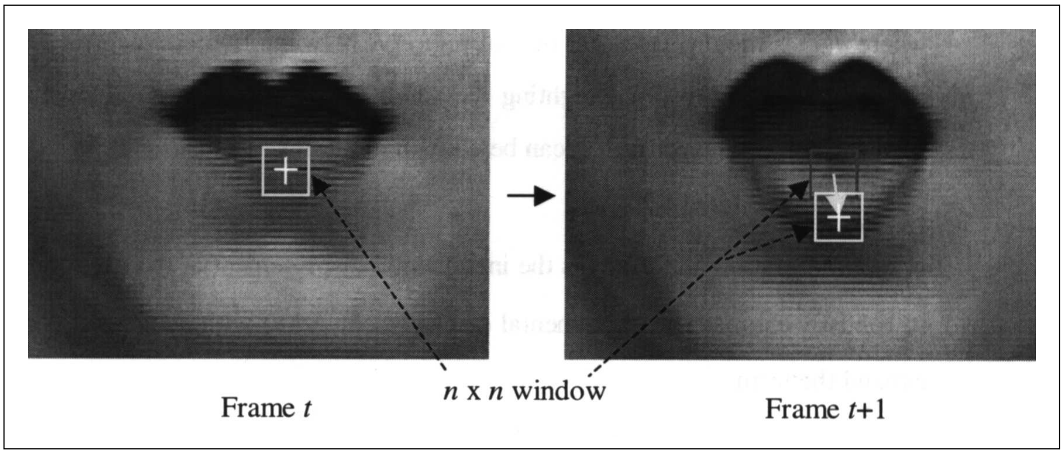


Figure 15: *Feature point tracking by Lien and Cohn* [64]. The actual feature point consist a $n \times n$ feature window that is tracked between any two consecutive frames (here, $n$ was chosen to be 13 pixels).

### 4.5.3 Spatial versus Spatio-Temporal Classification

The question whether spatial or spatio-temporal classification is better depends on the envisaged application. If we would like to achieve a direct mapping of facial expressions towards mental categories, spatio-temporal mapping may be more suitable. For example, a distinction between feigned and true smiles can be more easily achieved, if the dynamics of a given facial expression display is known. Traditional approaches for modeling temporal characteristics of facial motion have relied on hand-crafted rules [71][108][17]. Hence, human expertise for defining specific emotional characteristics is necessary. Real facial signals on the other hand consist of numerous distinct expression with specific facial action intensity evolutions. In addition, individual realizations of facial expressions often differ only in subtle ways. This makes the task of manually creating facial expression or emotion classes rather difficult. Spatio-temporal information may therefore only complicate facial expression analysis and classification stages, but be important when interpreting facial expressions. This is why an intermediate representation of facial expressions prior to interpretation is convenient, as it allows to represent facial actions with fewer parameters and thus allows for a better control. FACS is a suitable representative of an intermediate facial expression dictionary, as it is independent of mental displays and codes facial actions on a frame-per-frame base.

## 4.6 Facial Expression Intensity Measurement

It is not only important to locate facial activities, but intensities should be estimated as well as these allow to analyze the temporal evolution of a given facial expression display. Unfortunately, this important issue has been neglected too much in the literature.

*Lien* [64][61] estimated facial expression intensities using three different motion-based methods that served also for the detection of facial activities: feature point tracking, dense flow tracking and high gradient component detection. Since the displacement of facial feature points indicates changes of muscle contraction, it is possible to quantify expression intensities based on the motion of feature points. Expression estimation was done using a nonlinear mapping of the feature point displacement. Figure 15 shows a sample feature point and its motion between two subsequent frames. With the dense optical flow approach, expression intensities of individual training sequences frames were determined by experts prior to automated intensity analysis. Intensities where measured with a scale ranging from the neural expression with intensity 0.0 to the peak expression with intensity 1.0. Each training frame was attributed a weight vector, obtained after a PCA compression of the optical

| **Extraction Method** | **Sub-Type** | **Nb. of Components** | **Recognition Performance** |
|---|---|---|---|
| **Optical Flow** | Gradient | - | 55.8% ± 4.7% |
| | Correlation | - | **85.6% ± 3.3%** |
| | Smoothed | - | 53.1% ± 4.7% |
| **Difference Image** | PCA | 30 | 79.3% ± 3.9% |
| | FLD | 5 | 75.7% ± 4.1% |
| | ICA | 75 | **95.5% ± 2.0%** |
| | LFA | 155 kernels | 81.1% ± 3.7% |
| **Local Analysis**[1] | Gaussian Kernel | 15 x 15 kernel | 70.3% ± 4.3% |
| | PCA Random | 2 to 30 | 73.4% ± 4.2% |
| | PCA Fixed | 10 | 78.3% ± 3.9% |
| | PCA Jets | 2 to 17 | 72.1% ± 4.2% |
| | Gabor Jets | - | **95.5% ± 2.0%** |
| **Human** | Naive | - | 77.9% ± 2.5% |
| | Expert | - | **94.1% ± 2.1%** |

Table 8: *Extraction Method comparison by Bartlett* [4], tests were performed on a database with 12 different facial actions displays. [1]Local means here that the input image is divided up into mosaic-style patches that are processed independently. This approach shouldn't be confused with facial feature processing though, where the focus is on local facial features.

flow field. The intensity of a given test frame was estimated by comparing the test weight vectors with those obtained in the training sequences by applying the minimum distance criterion. Finally, furrow expression density was estimated with the squared difference criterion on multi-dimensional mean-variance vectors for facial expressions in the mouth and the front regions.

*Hong et al.* [51] used personalized galleries that feature face images with slightly different degrees of intensities. Totally four face images were recorded per person and expression. The whole-face template that matched best to a given test face determined both the prevalent facial expression and its intensity.

*Lisetti and Rumelhart* [65] did facial expression measurements using a neural network that focused both on the whole face, eyes, lids, and mouth. Facial expression intensities were computed only for neutral and smiling faces. The expression range encompassed either six or eight intensity levels. Good results were obtained also for previously unseen faces. Zooming into particular areas of the face improved facial expression intensity estimation compared to the results obtained by processing the whole face at a uniform scale.

*Kaiser and Wehrle's* [128] marker-based system was not only able to automatically code facial expression intensities according to FACS, but allowed to detect asymmetries as well. Hence a extension of the FACS coding system was proposed.[14]

## 4.7 Comparison of Selected Analysis Methods

*Bartlett et al.* [3][6][5][4] studied various spatial image analysis methods, among these component analysis methods such as Principal Component analysis, Fisher's Linear Discriminant, Independent Component Analysis, Local Feature Analysis, Gaussian kernels and Gabor wavelets. Further on, different optical flow methods as well as geometric feature measurement techniques were applied. Table 8 shows how these methods performed on the same database and under the same conditions. The correlation based optical flow performed best in the motion analysis category. Performance

[14]FEAT coding examples: AU12 cb: action unit 12 weak on the right hand side, left hand side hardly perceptible. AU 6 -c: action unit 6 unilateral, right hand side with hardly perceptible intensity.

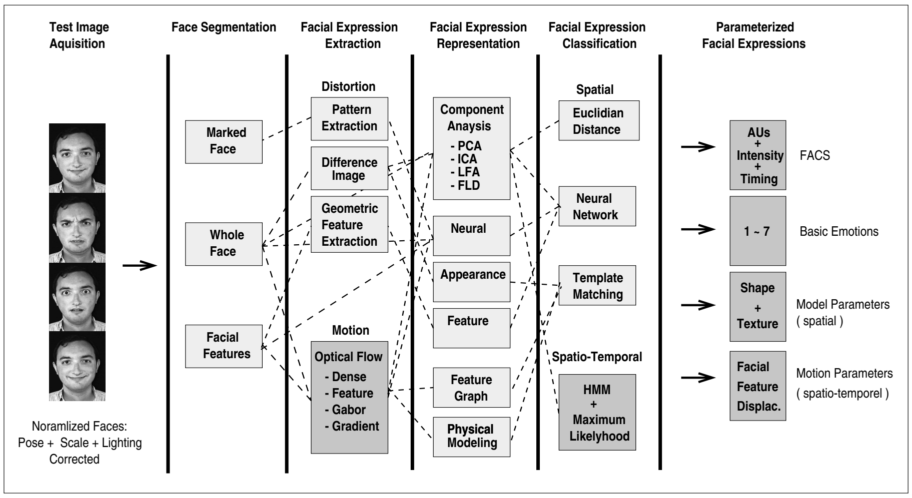


Figure 16: Canonical representation of facial expression analysis methods.

differences were less pronounced for the distortion based approaches. ICA and the Gabor based methods performed equally, but much better than the optical flow based methods.

# 5 Facial Expression Recognition Systems

In this chapter follows a review of a few representative facial expression analysis systems. First, we introduce a *marker-based* facial expression analysis system that allows FACS coding with a performance comparable to a human expert. Then we discuss *hybrid* approaches which employ several image analysis methods that complete each other and allow for a better overall performance. *Multi-modal* methods on the other hand integrate other non-verbal communication channels such as emotion in voice for improved interpretation results. Finally, *complete frameworks* feature multiple face processing strategies, combining different face processing task such as face recognition and facial expression analysis.

In order to facilitate the comparison of the facial expressions analysis systems discussed further below, various image analysis methods were represented in canonical frameworks, ordered according to the different face processing stages that were introduced in the previous chapter. Figure 16 shows a canonical facial expression analysis system.

## 5.1 Marker-based Systems

*Kaiser and Wehrle* [128][132] developed a system called FEAT (Facial Expression Analysis Tool). It uses markers[15] and was able to classify all FACS AUs reliably using a compiled neural network, see Figure 17. FEAT is a connectionist expert system that uses fuzzy rules, acquired from a FACS expert, to automatically measure facial expressions. This FACS expertise was transformed into a network structure by a compiler program. The resulting network is then able to do the classification task, using FACS as a coding language, see Figure 18. The FACS coding performance of FEAT is comparable to a human expert. The combination of a fuzzy rule base and an artificial neural network has proven to be appropriate when dealing with the multiplicity and complexity of facial behavior. The overall agreement between automated and human expert scorings was found to be 0.73, which exceeds the FACS Final Test Criterion of 0.70 [123]. However the system requires precisely positioned artificial markers and is therefore limited in its use to special experimental setups. Therefore, video

[15]Small plastic dots affixed to predefined locations of the subject's face. These allow to make tissue motion visible.

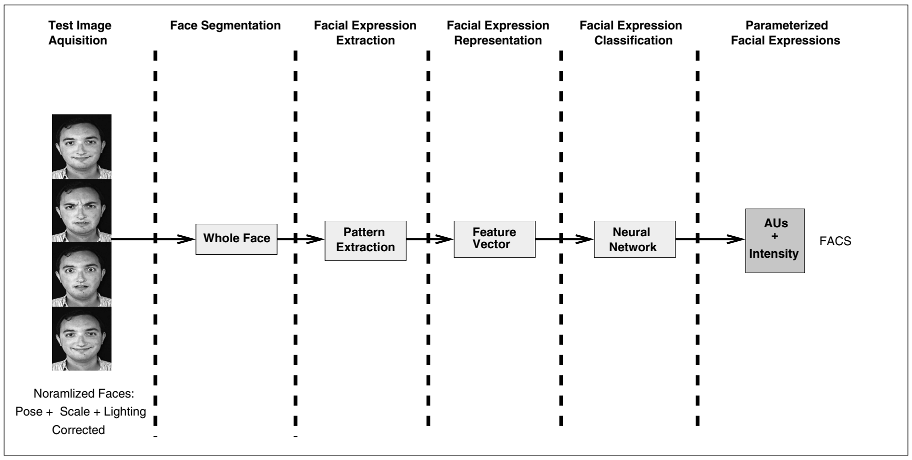


Figure 17: Canonical representation of a marker-based system by *Kaiser et al.* [128].

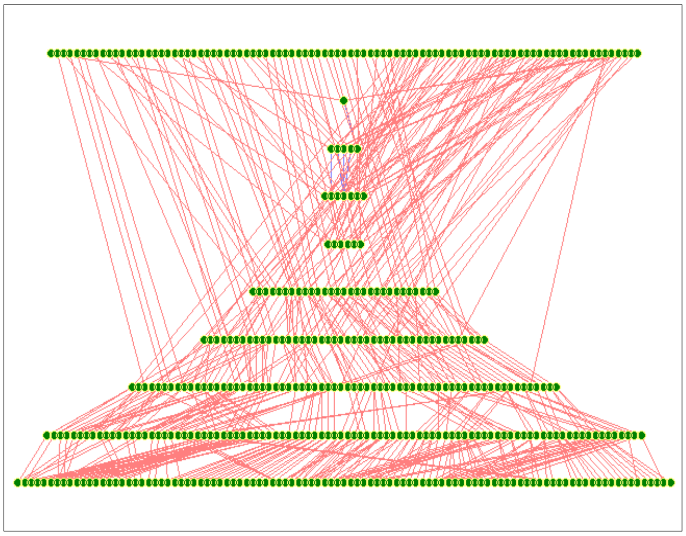

Figure 18: *Marker-based approach:* Compiled Neural Network Classifier by *Kaiser et al.* [128].

indexing would not be feasible due to the lack of marked faces. Kaiser and Wehrle explored also traditional classification approaches based on inference engines. These methods failed and did not produce satisfying results, as the rules found proved to be too vague [128]. Although most people recognize facial expressions more or less correctly, this recognition process is based on tacit knowledge. It is therefore difficult to formulate logical rules. Even with FACS, that allows to identify every single Action Unit, the coding of all possible combinations of AUs is rather complex. Another alternative would be the use of classical connectionist systems, e.g. neural networks trained with a back-propagation approach. Kaiser and Wehrle pointed out that this was not feasible, given the large training set that would have been necessary for a good classification of the many AUs and AU combinations [133]. A critical problem of the latter approach was also the poor ability to generalize over several individuals. Kaiser and Wehrle report that they were able to use these nets to classify the facial behavior of a certain person. A person dependent training set had to be created manually though. This made the advantage of the automation obsolete. With FEAT, a probabilistic approach that encompasses a language for expressing fuzzy rules together with an artificial neural network, it was possible to make the classification mechanism independent of individual physiognomy [128].

## 5.2 Hybrid Systems

We have seen above that every image analysis method has its own characteristics and focuses on different information contained in the test images. Hybrid systems employ several different image analysis methods in parallel and can thus combine extracted information. This might result in a more accurate representation of facial actions. We present two hybrid facial expression analysis systems that use both the FACS coding scheme and can therefore be compared on a common bases.

*Bartlett* [4] built a hybrid facial expression analysis system combining holistic ICA (PCA), geometric feature analysis and correlation-based optical flow into a single framework. These three methods were compared regarding to their contribution to the facial expressions recognition task [6]. This way one obtains some insight of how the system benefits from different analysis methods by looking at the correlations of the recognition performance. A combination of several facial expression analysis methods is most beneficial when the individual estimators produce very different error patterns. Figure 19 shows some correlations between individual analysis methods. Holistic PCA and optical flow are correlated with a coefficient of 0.52. The geometric feature-based method however has a more independent error pattern compared to both the optical flow and PCA approaches. Even though the stand-alone performance of the geometric feature-based system is low, it contributes to a hybrid system because it provides estimates that are independent from the other analysis methods. Bartlett estimated that without geometric feature measurement there would be a 40% decrease of the improvement gained by the combination of the tree analysis methods. It seems that the additional information yielded by the geometric feature measurement method was lost in the holistic methods either with the computation of the difference images, due to insufficient image resolution or suppressed by taking only those PCA components that were accompanied with important eigenvalues.

Bartlett's hybrid facial expression analysis system was able to code 12 FACS action units. Classification was done by combining the output of all three analysis methods into a single feed-forward neural network made up of 10 hidden and 6 output units. A winner takes it all (WTA) judgment approach was chosen to select the final AU candidate. The input of the neural network consisted of 50 component projections (PCA), 5 feature measures (geometric feature measurement) and 6 template matches (optical flow). Figure 20 shows the canonical representation of the Bartlett's hybrid system.

Interesting is the holistic spatial analysis approach that employs *difference images* followed by PCA compression and classification by a neural network. The PCA stage improved the recognition rate from 84% to 88.6%. Best performance was obtained using 30 component projections, respectively 75 for the ICA representation. Note that in contrast to most other motion analysis approaches, Bartlett used a neutral face frame also for the motion field analysis via optical flow. With other words, each test frame was not compared to another test frame previously recorded in a sequence, but to a neutral

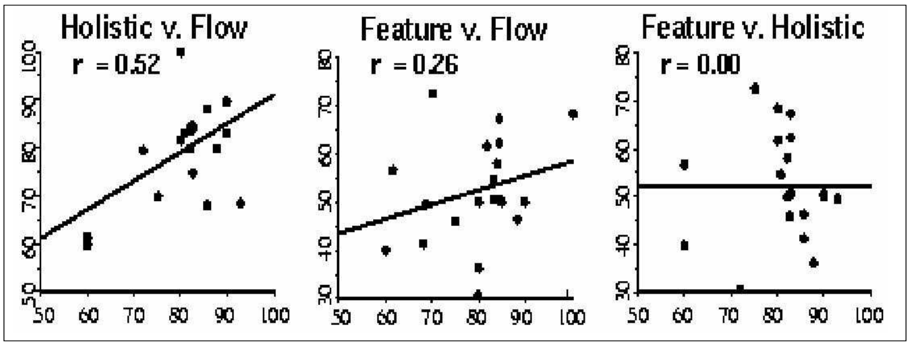


Figure 19: *Holistic PCA versus flow and geometric feature measurement methods.* The axis indicate recognition rates in percentage. Source: *Bartlett et al.* [6].

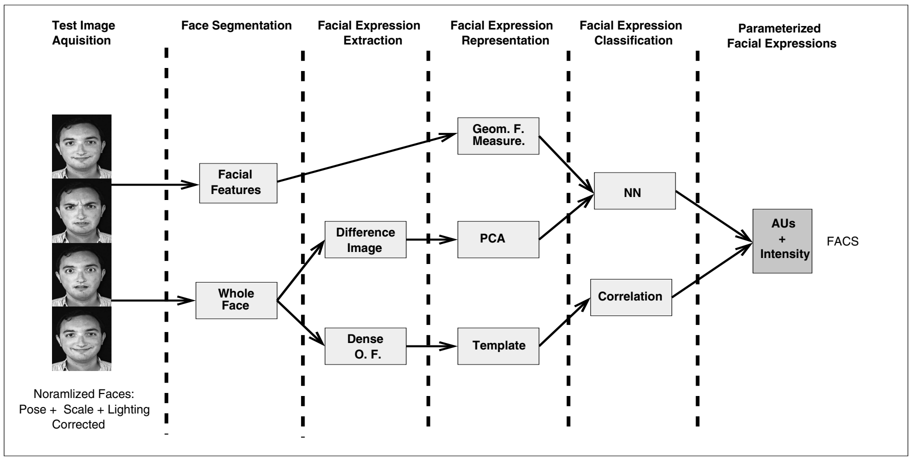


Figure 20: Canonical representation of the hybrid facial expression analysis system proposed by *Bartlett et al.* [4]. Note that not only PCA was used as a holistic representation method, but also ICA.

Figure 21: *The hybrid system of Lien* [64] is based on three extraction methods: On the left-hand side is shown the Dense Optical Flow, in the middle the Feature-Point Optical Flow and on the right-hand side the high gradient component detection approach.

face frame in the beginning of a given facial expression test sequence. Even though Bartlett's system is one of the most advanced systems today, it has also some drawbacks: first of all, the image alignment is limited to translations and scaling which is insufficient to align rotated face images. Alignment need manual intervention: face positions are determined by clicking the eye centers in the first face frame of a test sequence. Second, only rather limited image data sets can be used, as the first frame always has to contain a neutral face for the creation of difference images. In addition, tests were performed on image sequences containing 6 frames from low to high expression intensity. These were not sampled at full frame rate but frames had to be dropped in between to allow for a full range of expressions intensities. In other words, manual time-warping was forced onto the frame sequences. Third, FACS coding comprises also intensity coding and the determination of onset, apex and offset, features that were not implemented for the system described above. Images were mostly captured from experts performing AU unit and AU pair combinations rather than more complex AU combinations with three or more AUs involved from both the lower and upper face parts displayed at the same time.

*Lien* [62][63][60][61][64] *and Cohn et al.* [25][26] introduced a system, that is based on three face analysis methods: Holistic dense-flow tracking, feature point tracking and high-gradient component detection, see Figure 21. Facial feature point tracking was combined with a pyramid method, a multiscale approach which is especially sensitive to subtle feature motion and allows to track large displacement of feature motion with sub-pixel accuracy. Lien placed feature points into facial feature regions such as the brows, eyes, nose and mouth. The feature point tracking approach applied is based on the motion tracking of $n \times n$ feature windows between consecutive frames, where a value of 13 pixels was chosen for $n$. The dense optical flow technique employed by Lien is based on *Wu's* [103] approach and allows the estimation of entire facial motions in an accurate way. Forehead, cheek and chin regions do carry important expression information, especially wrinkles. Facial motion associated with wrinkles produces gray value changes in grayscale face images. These can be extracted by using high gradient component motion analysis based on horizontal, vertical and diagonal line or edge detectors in the spatial and frame comparisons in the temporal domain. Figure 22 shows the facial expression analysis system implemented by Lien and Cohn. The *whole* facial expression analysis framework is situated in the spatio-temporal domain. Spatio-temporal classification was achieved with Hidden Markov Models (HMM).

Lien's facial expression analysis system performed well also on rather difficult image sequences. Test were also performed on baby face sequences. Baby faces are different from those of adults both in morphology and tissue texture. The image alignment procedure used in Lien's system is more sophisticated than the one found in Bartlett's system, but is neither completely automated, as three facial feature points need to be manually selected on the face in the first frame of each test

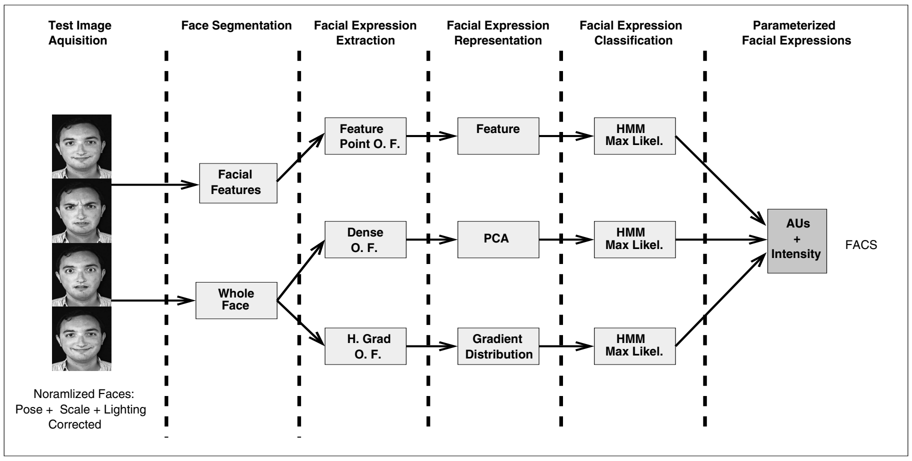


Figure 22: Canonical representation of the hybrid facial expression analysis system proposed by *Lien* [64].

image sequence, in order to define the position and scaling of the current test face and allow for warping of the latter if necessary. The system handles 15 AUs and AU combinations situated in the brow, eye and mouth areas. As we have seen further above, Lien's system is also able to measure facial expression intensities with the aid of three different methods. Since the whole facial expression analysis is performed in the spatio-temporal domain, *heavy computational requirements* arise especially with the dense optical flow approach. The latter is probably the most important analysis technique in the system's framework, as it tracks the face in a holistic way, measuring the displacement of all pixels. Lien reported that dense flow tracking would take up to hours for the analysis of a few frames. Another open question is the robustness of the whole system, as it is entirely based on optical flow methods. These are problematic with lighting variations and iconic changes.

## 5.3 Multimodal Approaches

Today, most facial expression analysis systems are of the *unimodal* type as they focus only on facial expressions for the determination of mental activities. As we have seen, there are other sources of non-verbal communication apart from facial expressions as well, probably the most important one being voice. At present, most attempts of channel fusion are of the *bimodal* type; these systems integrate voice as an additional channel to facial expressions. It may be difficult though to obtain coherent emotion expression in voice as there are various ways of how to express a certain emotion and in addition, these are heavily culture dependent.

*Cohn and Katz* [113] developed a prototype of a system that is based on vision and acoustics and is able to discriminate communicative intents in infant-directed and felt emotion in adult-directed speech. Vocal expressions are carried by prosodic features which include the fundamental frequency, intensity and rhythm of the voice. Cohn and Katz focused on the fundamental frequency as it may be the most important voice feature for emotion detection. They did a quantitative modeling of the fundamental frequency contours and reported that in the study with facial and prosodic measures, convergence with self-reported emotion was greatest, when both types of measures were combined. Recognition performance for infant-directed speech was higher due to exaggerated prosody which can be discriminated better than prosody in adult-directed speech. The systems applied for speech and vision analysis were separated and a concordance of prosodic measures with facial expressions was performed by assessing facial expressions with surface electromyography (EMG). Better results could probably be achieved if one would rely not only on the fundamental frequency for emotion recognition but instead using prosody as well. *Chen et al.* [23] proposed a different way of combining the two modalities, using a rule-based approach that is applied directly on the audio and vision signals to combine them prior to classification instead of doing independent classification of both the video and the audio paths. A completely different approach took *Kaiser et al.* [130] tackling the difficult task of *context* information integration into a facial expression analysis system. This was done using an interactive computer game, where the player is involved and not only observer. All data of the subject's manipulations and interventions were recoded continuously. This context information is useful for the interpretation of simultaneously recorded facial expression behavior. Interpretation in this case was done by humans.

The evaluation of multiple communication channels done in parallel may foster robustness as well as correct interpretation of facial expressions in ambiguous situations. The latter occur when one channel contradicts the other while dominating and thus giving the message conveyed by the facial display another meaning. An example is smiling occurring at the same time with a menacing timbre in the voice.

## 5.4 Complete Frameworks

Currently most systems described in the literature are either specialized for face detection, face recognition, gender detection, age estimation, ethnicity determination or facial expression recognition. Systems that would combine all of these face processing tasks may allow for fewer manual interventions and more information to be extracted from faces. Complex face processing tasks such as facial expression recognition may well benefit from parameters obtained by age estimation, gender detection and ethnicity determination, because all of these characteristics introduce both spatial and temporal face variations. Unification of several face processing tasks into a single framework promises to deliver results that go beyond to what could be expected by a single approach only, due to *synergy* effects.

Appearance-based models use both shape and texture models and allow for multiple face processing tasks such as face recognition and facial expression analysis. Furthermore, appearance-based models have the interesting property of analyzing faces both in a spatially local (shape models) and holistic way (texture models), allowing for the tracking of both facial features and holistic texture changes in skin caused by furrows and wrinkles. Appearance-based models work also on 2D and 3D rotated faces if the whole face is visible. *Lanitis et al.* [59] was able to extract facial appearance parameters that contained 3D pose estimations, identity, gender and facial expression and face synthesis information. The applied facial analysis approach can be divided into two main phases. First, flexible models of facial appearance are generated. Shape models are derived from a set of training images by statistical analysis of the positions of landmark points. These represent main facial features and are placed manually on training images prior to the shape model creation process. In a second phase the models are used for coding and interpreting of the face images. Facial features are located in every new test image using ASM search [28] based on the flexible shape models obtained during training. The automatically located model points are then transformed into shape model parameters. Furthermore, graylevel information is collected at each model point and transformed into local profile model parameters. Then, the test face is deformed to the mean face shape and the graylevel shape-free model is extracted. Figure 23 illustrates the second phase in canonical form. One of the main drawbacks of this approach is the manual intervention necessary for the generation of the appearance-based models during the training phase. An open issue is the probably huge number of models (both shape and texture models) needed for the classification of facial intensity levels.

*Cottrell and Metcalfe* [30] proposed a somewhat complete approach using a neural network based framework that allowed for gender detection, face and expression recognition. The face images of 20 people performing 8 different expressions were tested, making a total of 160 stimuli. In a first stage, these images were compressed with a three layer network ($4096 \times 40 \times 4096$). The representation formed in the hidden units was then used as input for a 2 layer network trained for gender and identity classification and a 3 layer emotion network ($40 \times 20 \times 8$) trained to classify faces according to facial expression, see Figure 24. Results showed that the "gender and identity network" was able to identify 99% of the training set and to perfectly classify the test faces according to their gender.

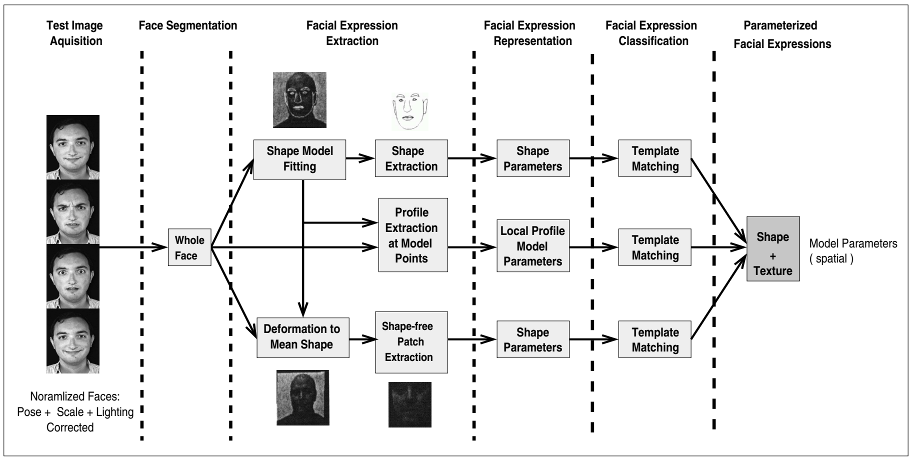


Figure 23: Canonical representation of a system based on appearance models by *Lanitis* [59].

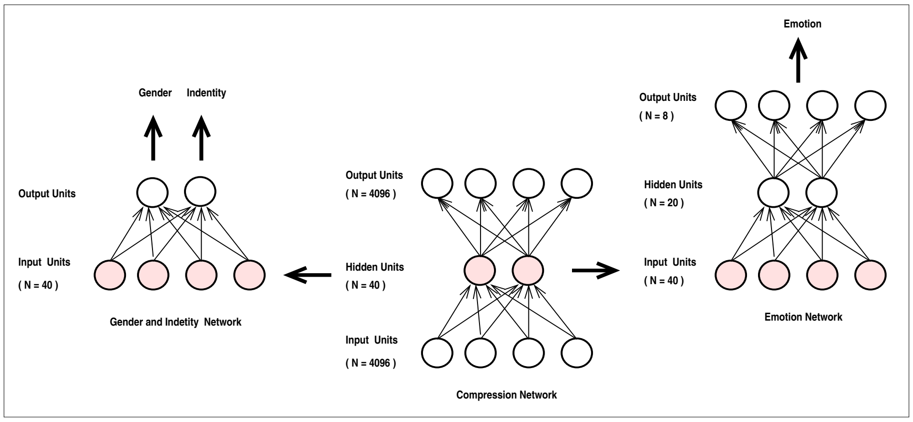


Figure 24: Complete Framework: Neural networks for different face processing tasks by *Cottrell and Metcalfe* [30].

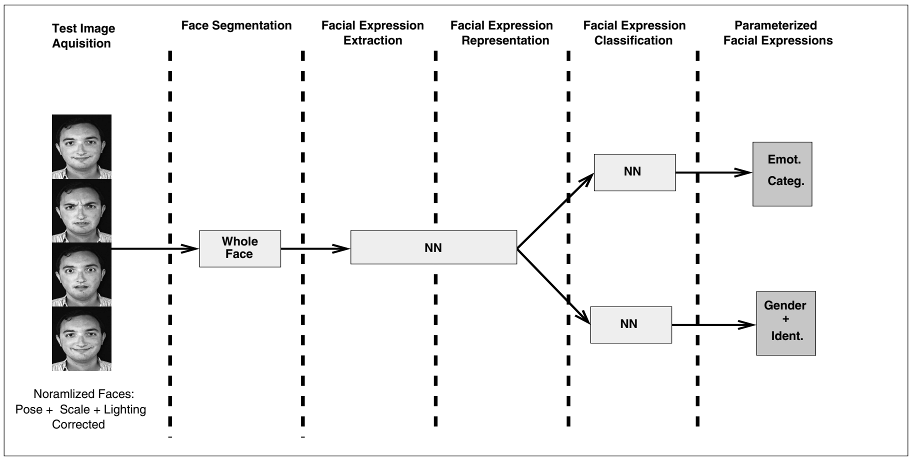


Figure 25: Canonical representation of a neural network based approach by *Cottrell and Metcalfe* [30].

The "emotion network" was able to distinguish some of the positive emotions (astonished, delighted, pleased, relaxed), but was unable to reliably classify the negative ones (sleepy, bored, miserable, angry). *Valentin et al.* [99] argued that this could be due to the fact that people are better at displaying positive emotions on command, than negative ones, and therefore feigned negative expressions might be more similar to each other than positive ones. The above described "emotion network" seems to be rather small, especially with regard to the number of hidden units and thus it may not allow for much discrimination capacity. An open question with this neural network approach remains training. As long as we attempt to classify emotion categories directly, it may be less an issue, but if we would attempt to use the FACS coding scheme, this would become a problem due to the number of different training samples needed to distinguish all AUs as well as their combinations. Figure 25 finally shows the canonical representation of the neural network approach proposed by Cottrell and Metcalfe.

## 5.5 Comparison of Selected Systems

It is difficult to compare existing facial expression analysis systems due to differences in the databases used both for training and testing. Especially important is the number of subjects and facial expression displays per subject contained in the test databases. Furthermore, different facial expression classification strategies were pursued. The systems either attempted to separate emotion classes or were based on FACS coding. Some systems tried also to distinguish between facial expression intensities. For a true comparison, systems would have to be tested on the same data and code facial expressions in the very same way. Nonetheless, we tried to summarize results obtained with the most prominent facial expression analysis systems found in the literature, see Table 9. *Bartlett* [4], *Lien* [64] and *Cohn* [25] used the FACS coding scheme with approximately the same AU discrimination capacity. Lien obtained slightly better results with its dense optical flow approach, whereas Bartlett's difference image based holistic approaches lead to a better recognition performance. Note also that in the latter approach optical flow was applied between a neutral frame and a given test image and not between consecutive frames as in the other approaches. *Lanitis et al.'s* [59] system delivered a lower facial expression recognition performance than the other systems. Probably this doesn't mean that their methods are inferior in comparison to those of other approaches, but rather that a difficult test set was used. Indeed, their database contained also difficult test images with partly occluded faces.

| **Authors** | **Nb. of Subjects**[2] | | **Nb. of Faces**[5] | | **Nb. of Classes** | **Analysis Method** | **Rec.**[1] **Rate** |
|---|---|---|---|---|---|---|---|
| | Test | Train | Test | Train | | | |
| Bartlett [4][6] | 20 | 20 | 240[10] | 1200[11] | **6**[3]+ **6**[4] | Gabor filters | **96%** |
| | 20 | 20 | 240[10] | 1200[11] | **6**[3]+ **6**[4] | ICA | **96%** |
| | 20 | 20 | 240[10] | 1200[11] | **6**[3]+ **6**[4] | Optical flow | **86%** |
| Mase [71] | 1 | N/A | 1 | N/A | **4** | Optical flow | **86%** |
| Yacoob et al. [105] | 32 | 32 | N/A | N/A | **7** | Optical flow | **87%** |
| Kobayashi et al. [57] | 30 | N/A | N/A | N/A | **6** | Feedforw. NN | **90%** |
| Rosenblum et al. [86] | 32 | 20 / 14 | 34 sequences | 20 / 14 sequences | **2** | Radial NN, Optical flow | **88%** |
| Padgett et al. [78] | 12 | N/A | N/A | N/A | **6**[7] | PCA | **86%** |
| Eessa et al. [40] | 8 | 8 | 8 sequences | 8 sequences | **6** | Param. models of Optical flow | **98%** |
| Lanitis et al. [59] | 30 | 30 | 300 | 390 | **7** | Appearance-based models | **74%** |
| Cohn et al. [25] | 30 | N/A | N/A | N/A | **15**[6] | Feature Point Optical Flow | **86%** |
| Lien [64] [6] [9] | 25 | N/A[12] | 75*20[8] | 60*20[8] | **3**[3] | Feature Point | **85%** |
| | 25 | N/A[12] | 150*20[8] | 120*20[8] | **6**[4] | Optical Flow | **88%** |
| | 15/22/23[13] | N/A[12] | 60*20[8] | 45*20[8] | **3**[3] | Dense Optical | **92%** |
| | 15 | N/A[12] | 90*20[8] | 60*20[8] | **6**[4] | Flow | **92%** |
| | 30/50[13] | N/A[12] | 160*20[8] | 100*20[8] | **3**[3] | Grad. Optical | **85%** |
| | 25/50[13] | N/A[12] | 150*20[8] | 120*20[8] | **6**[4] | Flow | **81%** |

Table 9: *Summary of selected automated facial expression analysis systems.* [1]These models were tested on different data sets with different levels of difficulty and therefore the obtained performances should be compared carefully. Classification categories were feigned expressions of emotion, except where otherwise indicated. [2]Number of faces of distinct persons in the database. [3]Upper-face part FACS coding. [4]Lower-face part FACS coding. [5] Either frames or image sequences. [6]FACS classified facial actions. [7]Classified images from pictures of facial affect [122] in which trained actors performed the muscle contractions empirically associated with certain emotional states. [8]Image sequences were of arbitrary length varying from 9 to 47 frames. The average number of images per expression sequence was about 20. [9]Note that the number of image sequences used for dense optical flow is a subset of the image sequences used for the feature point tracking analysis, which in turn is a subset of those used in the high gradient component analysis study. This is true for both training and testing sequences. [10]Only one (medium) intensity level out of three was chosen for the recognition performance test. Totally 20 subjects performing 12 different facial action at one intensity level. [11]20 subjects performing 12 different facial actions over 5 intensity levels. [12]The whole training and test database contained 94 subjects of different ethnicity, gender and age. [13]Different number of participating subjects per AU display.

# 6 Discussion

In this chapter, we are going to discuss the difficulties that arise, when attempting to compare different facial expression analysis systems. Then, we point out some shortcomings of current approaches and suggest some ideas for future work in the field of facial expression analysis and recognition.

## 6.1 Common Ground for Comparison

It is difficult to compare current facial expression analysis systems because of different implementation decisions that were taken, respectively different test beds that were used for the assessment. First of all, there are systems that are conceived for the analysis of still images, whereas others take into account temporal information and hence analyze image sequences. This makes it virtually impossible to compare obtained facial expression recognition results. Secondly, facial expression were *not coded in a uniform way*. Only a few system employ the FACS coding scheme, while most systems attempt to map extracted facial expressions directly to basic emotions. Thirdly, almost all systems were tested on *different databases*. There is a lack of an all encompassing and common facial expression database that would allow for testing facial expression analysis methods in a more transparent way. Such a database should feature both training and a test sets containing entries with following characteristics:

- Faces of different **age**, featuring both adults and children. Baby faces have smoother skin than those of adults and contain less wrinkles. Further on, the facial anatomy is different, e.g. cheeks are more prominent. In contrast, old people tend to have more intransient wrinkles, which could be a potential problem for facial expression extraction algorithms.
- Subjects with different **gender** and **ethnicity** and **cultural background**. Facial expressions are culturally dependent, both in intensity, winks and controlling of the facial display.
- Various different **poses** with sidewards tilted and both in-plane and out-of-plane rotated faces. Especially out-of-plane rotated faces pose great problems for automatic facial expression analysis systems as facial features may be distorted, only partially visible or completely out of sight.
- **Moving and talking heads** with emotional voice as an important additional non-verbal channel.
- **Bearded faces**, partially **occluded faces** as well as faces with **glasses**.
- Changing **ambient conditions** with lighting variations and cluttered scenes so that the overall robustness could be tested.
- Longer **image sequences** would allow to test system stability, sequences with feigned and true emotional display. Since a facial expression system should also be able to detect facial expressions in static scenes, independent **still images** would be needed as well.
- A **multitude of facial displays**, not only basic emotions, but more subtle ones such as boredom and miserable, and also faces showing cognition processes, social interactions and psychophysiological signals.
- The database should be **balanced**, meaning that all subjects would display all facial expressions.

Following predicative measurements are important in the context of facial expression analysis:

- **Generalization performance** measurements that allow to test how a given system handles novel faces.
- Comparisons to facial expression recognition performance obtained by human experts or systems that have passed the FACS coding test issued by Ekman et al., in order to provide **ground truth**.
- Performance comparisons on **specific facial expressions** are important in order find out where a given facial expression extraction or classification method fails.

## 6.2 Shortcomings of Current Approaches

Most systems attempt to map facial expressions directly to basic emotions. The latter have the advantage of being independent of cultural backgrounds, but as highlighted before, emotions are only one source for facial expression display and therefore such a categorization is too coarse. Currently, there are only three systems that use the FACS coding scheme [4][64][128]. The latter achieves expert performance but is based on markers. The other two systems are marker-independent, but are able to code a *only subset* of all FACS AUs, AU combinations and AU intensities. This is not surprising, because FACS coding is a difficult task. About 7000 Action Unit combinations have been identified!

Most facial expression analysis systems *lack robustness* towards lighting and pose changes and are unable to analyze partly occluded faces. Many analysis methods make the hypothesis that the test face is centered in the image, seen from a near frontal view throughout the sequence and with small rigid head motion between any two consecutive frames, or with other words, spatially bounded.

In current facial expression analysis systems, too much *manual intervention* for accurate normalization is necessary: localization of facial feature points in the initial frame, facial feature template selection and manual warping of video sequences. Further on, there are only a few multimodal systems that take also acoustic signals into account. Although facial expression often occur during conversations [116], none of the cited approaches did consider this possibility.

## 6.3 Future Directions

Apparently, human beings are able to judge facial expressions on previously unseen faces, we could therefore assume that the brain stores some kind of generic neutral face models. Learned Generic face models for independent facial expression extraction would allow the analysis of faces, where no neutral face is available for comparison. Generic neutral faces could be obtained by averaging a multitude of neutral faces .

In most facial expression analysis systems, a given test face is warped or rotated into a frontal-face position prior to the analysis of facial actions. Alternatively, instead of attempting to normalize pose changes prior to facial expression analysis, one could analyze facial expressions directly, where facial expressions would be treated differently with regard to the viewing angles. This would address the problem of *perspectively distorted faces*. Connectionist models use a distributed rather than localized way of representing faces. Faces share the same storage space, and hence, representations of similar face can interfere with each other [99]. This could be a particularly suitable approach for representing the face from different angles. Another advantage of neural networks is that they preserve the relationship between features and texture which is important for face analysis. Information used in connectionist models is automatically derived from the statistical structure of the faces, and therefore, the difficult problem of selecting individual features is avoided. This for example has to be done manually in the case of appearance-based models, where the shape model is learned by indicating feature points manually.

So far we have focused on facial expression analysis and recognition, respectively the classification of facial expressions and their coding into action units. The ultimate goal would be the mapping of observed facial expressions into mental activity states. A mental activity analysis system could use FACS codes obtained by an automatic facial expression analysis system and map these into mental activity categories using a dictionary and rules. Evidence that this can be achieved is supported empirically: FACS scores yield highly accurate pre- and postdictions of the emotions signaled to observers in more than fifteen cultures, western and non-western, literate and preliterate [119]. An example of an emotion mapping dictionary are the FACS emotion prediction tables [123].

In face recognition intra-personal variations such as facial expressions shouldn't influence recognition results, whereas in facial expression analysis we attempt to gain a maximum of stability with regard to inter-personal variations, with a great sensibility to intra-person variations. This doesn't mean though that facial expression analysis is totally orthogonal to face recognition. On the contrary, for the task of facial expression analysis, it could also be an asset to have access to the identity of the

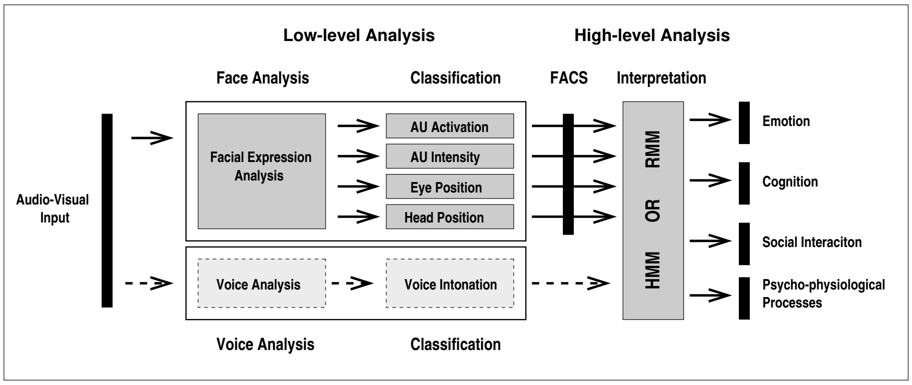


Figure 26: Mental Activity Analysis Framework.

test subjects. Since every human being displays facial expressions with proper intensity and dynamics, identity would allow for fine-tuning where the use of personalized facial expression galleries could allow for a better mapping of facial expression towards mental states. This intra-personal variations may be learned in an adaptive manner. *Hong et al.* [51] utilized *elastic graph matching* for both face recognition and facial expression recognition based on personalized galleries. For an incoming test face with a specific expression, first one tries to find the best matching person whose face is stored in a so called gallery. It contains a set of face images per subject, showing different facial expressions. Then the personalized gallery is used to identify the facial expression on the test face under the assumption, that two persons that look similar would also display facial expressions in a similar way.

HMMs or RNNs could cope with the temporal aspects of mental activities and would probably be better suited than static classifiers, especially for interpretation. Figure 26 shows the outline of a bimodal mental activity analysis system that uses voice as an additional channel. There are still open questions when it comes to the discrimination of mental states: how important context information and social interaction such as winks are for the mapping of mental states. Taking these into account may involve *reasoning* and therefore an implementation to analyze this additional information would hardly be feasible. Another question is how in humans facial expression analysis interferes with the short-time memory, respectively how long perceived facial expressions are retained and compared to previously perceived expressions. This important when attempting to analyze and model temporal aspects of emotions (e.g. feigned versus true smiles). Furthermore there is yet another type of facial expressions not discussed so far*: facial color* changes caused by emotions such as shame (blushing) and fear (blanching). These displays are difficult to perceive, also because there is a great deal of variation in human skin color. It is difficult to obtain training samples of people displaying these kind of signals, as they cannot be feigned easily.

# 7 Conclusions

While humans do facial expression analysis, recognition and interpretation effortlessly, these are very complex tasks, as there is a large variance between individual displays of facial expressions and it is not easy to measure facial behavior objectively.

Automated facial expression analysis has become an important research area in computer vision. Promising applications such as advanced human machine interfaces could alter significantly the way we communicate with computers. Today, most systems attempt to map facial expressions directly into basic emotional categories, which is a limited approach because emotions occur in blends and each basic emotion can be divided into a variety of sub-categories. Also, emotions are not the only cause for facial expressions. Systems not relying on artificial highlighting have still difficulties with robustness and no system achieved perfect FACS coding. Currently, most promising seem to be *hybrid* *approaches,* that focus both on holistic and local face analysis, as do for example the systems proposed by *Bartlett* [4] and *Lien* [64]. Both systems code facial expressions using the FACS coding scheme. Promising facial expression analysis methods are Gabor wavelet representations [4][109], local spatial analysis [4], optical flow based approaches using feature point tracking [64], spatially acting methods such as PCA or ICA [4], as well as appearance-based models that analyze both shape and texture of faces [59]. Few research has been done on systems that go beyond the scope of facial expressions analysis: *Multimodal systems* that address facial expression together with voice analysis could allow for better interpretation of facial expressions. *Complete frameworks* that address facial expression analysis, age, gender, ethnicity and identity in a unified manner are of great interest and may deliver superior results. Most current facial expression analysis systems are not yet robust and need manual intervention at one point or another, especially for pose normalization an face localization.

The scope of this survey was limited primarily to the computer science approach on facial expression analysis and recognition. Psychological and neurological aspects were mentioned briefly in order to describe facial characteristics and human behavior in the context of facial expressions, respectively gaining some insight on how human brains processes facial expressions. Finally, we have seen, that it is important to accurately represent facial expressions with a dictionary such as FACS. Facial expression interpretation would be the next challenging step once facial expression could reliably recognized.

# 8 Acknowledgments

The authors would like to thank Susanne Kaiser and Thomas Wehrle from the University of Geneva for making available their database as well as fruitful discussions concerning FACS in general. Further thanks go to the SNSF (Swiss National Science Foundation) for funding this project under project number 21-54000.98.

# 9 Literature

The following reference section is divided into the four main research and application fields where facial expression analysis and recognition is of importance, namely computer and engineering sciences, psychology, neuroscience and medicine.

## 9.1 Computer Science

[1] M. Abdel-Mottaleb, R. Chellappa, and A. Rosenfeld. Binocular motion stereo using map estimation. *IEEE CVPR*, pages 321–327, 1993.

[2] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. In *IJCV*, volume 12, pages 43–77, 1994.

[3] Marian Stewart Bartlett. Unsupervised learning of invariant representations of faces through temporal association. In *The neurobiology of computation: Proceedings of the Annual Computational Neuroscience Meeting*, 1994.

[4] Marian Stewart Bartlett. *Face Image Analysis by Unsupervised Learning and Redundancy Reduction.* Phd thesis, University of California, San Diego, 1998.

[5] Marian Stewart Bartlett, H. Martin Lades, and Terrence J. Sejnowski. Independent component representations for face recognition. Technical report, University of California and the Salk Institute, La Jolla, 1998.

[6] Marian Stewart Bartlett, Paul A. Viola, Terrence J. Sejnowski, Beatrice A. Golomb, Jan Larsen, Josep C. Hager, and Paul Ekman. Classifying facial action. In *Advances in Neural Information Processing Systems 8, MIT Press*, 1996.

[7] B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. Technical report, University of Oxford, Dept of Engineering Science, 1997.

[8] B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. In *Proceedings of Int. Conference on Computer Vision*, page (to appear), 1998.

[9] Sumit Basu, Alex Pentland, and Irfan Essa. Motion regularization for model-based head tracking. Tech Report 362, MIT Media Laboratory, Perceptual Computing Group, 1996.

[10] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[11] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[12] A. Bell and T. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[13] S. Ben-Yacoub. Fast object detection using mlp and fft. Tech. Report RR97-12, IDIAP, 1997.

[14] S. Ben-Yacoub, B. Fasel, and J. Luettin. Fast face detection using mlp and fft. In *Proc. Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, 1999.

[15] D. Beymer, A. Shashua, and T. Poggio. Example based image analysis and synthesis. AI Memo 1431, Massachusetts Institute of Technology, 1993.

[16] M. J. Black, D. J. Fleet, and Y. Yacoob. A framework for modeling appearance change in image sequences. In *Sixth International Conf. on Computer Cision, ICCV'98, IEEE Computer Society Press*, January 1998.

[17] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Fifth International Conf. on Computer Vision, IEEE Computer Society Press*, pages 374–381, July 1995.

[18] Michael J. Black and Yaser Yacoob. Tracking and recognizing facial expressions in image sequences, using local parameterized models of image motion. Technical report, Xerox Palo Alto Research Center, 1995.

[19] Michael J. Black and Yaser Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. Technical report, Xerox Palo Alto research center, 1995.

[20] A. Blake, M. Isard, and D. Reynard. Learning to track the visual motion of contours. *Artificial Intelligence*, 78:101–134, 1995.

[21] Roberto Brunelli and Tomaso Poggio. Face recognition: Features versus templates. In *IEEE Transcations on Pattern Analysis and Machine Intelligence*, volume 15, pages 1042–1052, October 1993.

[22] W. Cai and J. Wang. Adaptive multiresolution collocation methods for initial boundary value problems of nonlinear pdes. *Society for Industrial and Applied Mathematics*, 33(3):937–970, June 1996.

[23] Lawrence S. Chen, Thomas S. Huang, Tsutomu Miyasato, and Ryohei Nakatsu. Multimodal human emotion / expression recognition. In *Proceedings of the 2nd Int. Conf. on Automatic Face and Gesture Recognition*, Nara, Japan, April 14-16 1998. IEEE.

[24] Tanzeem Choudhury, Brian Clarkson, Tony Jebara, and Alex Pentland. Multimodal person recognition using unconstrained audio and video. Tech report, MIT Media Laboratory, Perceptual Computing Group, 1997.

[25] J. Cohn, A. Zlochower, J. Lien, Y.-T. Wu, and T. Kanade. Automated face coding: A computer-vision based method of facial expression analysis. *Psychophysiology*, 1998.

[26] Jeffrey F. Cohn, James Jenn jier Lien, Takeo Kanade, Wei Hua, and Adena Zlochower. Beyond prototypic expressions: Discriminating subtle changes in the face. In *Proceedings of the 7th IEEE Robot and Human Communication Workshop*, September 30 - October 2 1998.

[27] P. Comon. Independent component analysis - a new concept? *Signal Processing*, 36:287–314, 1994.

[28] T. F. Cootes, C. J. Taylor, and A. Lanitis. Multi-resolution search using active shape models. In *12th International Conference on Pattern Recognition*, volume 1, pages 610–612, Los Alamitos, California, 1994. IEEE CS Press.

[29] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. In *University of Manchester, Wolfson Image Analysis Anit, Department of Medical Biophysics*, 1998.

[30] G. W. Cottrell and J. Metcalfe. Empath: Face, gender and emotion recognition using holons. In R. P. Lippman, J. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems*, volume 3, pages 564–571, 1991.

[31] I. Craw and P. Cameron. Face recognition by computer. In David Hogg and Roger Boyle, editors, *British Machine Vision Conference*, pages 489–507. Springer Verlag, 1992.

[32] Matthew N. Dailey and Garrison W. Cottrell. Task and spatial frequency effects on face specialization. Technical report, U.C. San Diego, Department of Computer Science and Engineering, 1997.

[33] Trevor J. Darrell, Michael J. Black, and Yaser Yacoob. Task-specific gesture analysis in real-time using interpolated views. Tech Report No. 364, Mit Media Laboratory Perceptual Computing Group, 1995.

[34] J. Daugman. Complete discrete 2d gabor transform by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36:1169–1179, 1988.

[35] Benoit Duc. *Feature Design: Applications to Motion Analysis and Identity Verification*. These no 1741, EPFL, Switzerland, 1997.

[36] G.J. Edwards, C.J. Taylor, and T.F. Cootes. Learning to identify and track faces in image sequences. Technical report, University of Manchester, Wolfson Image Analysis Unit, Department of Medical Biophysics, 1997.

[37] Peter Eisert and Bernd Girod. Facial expression analysis for model-based coding of video sequences. In *Picture Coding Symposium*, pages 33–38, September 1997.

[38] Peter Eisert and Bernd Girod. Model-based estimation of facial expression parameters from image sequences. In *International Conference on Image Processing (ICIP 97), Santa Barbara*, volume 2, pages 418–421, October 1997.

[39] A. Irfan Essa and Alex Pentland. A vision system for observing and extracting facial action. Tech Report 247, MIT Media Labratory, Perceptual Computing Section, 1994.

[40] I. Essa and A. Pentland. Coding, analysis , interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machin Intelligence*, 19(7):757–763, 1997.

[41] I. A. Essa and A. P. Pentland. Facial expression recognition using a dynamic model and motion energy. In *In Proc. 5th Int. Conf. on Computer Vision, IEEE Computer Society Press*, pages 360–367, 1995.

[42] I.A. Essa. Analysis, interpretation and synthesis of facial expressions. In *Vismod*, 1994.

[43] I.A. Essa, T.J. Darrell, and A.P. Pentland. Tracking facial motion. In *Vismod*, 1994.

[44] Irfan Essa, Sumit Basu, Trevor Darrell, and Alex Pentland. Modeling, tracking and interactive animation of faces and heads using input from video. Technical Report TR 370, Mit Media Laboratory, Perceptual Computing Section, 1996.

[45] Irfan A. Essa and Alex Pentland. Tracking facial motion. Tech Report 272, Mit Media Laboratory, Perceptual Computing Section, 1994.

[46] Irfan A. Essa and Alex P. Pentland. Coding, analysis, interpretation and recognition of facial expressions. Tech Report No.325, Mit Media Laboratory, Perceptual Computing Section, 1995.

[47] B. Fasel, S. Ben-Yacoub, and J. Luettin. Fast multiscale face detection. Cummunication com98-04, IDIAP, 1998.

[48] R. Fisher. The use of multiple measures in taxonomic problems. *Ann. Eugenics*, 7:179–188, 1936.

[49] D.J. Fleet and A.D. Jepson. Computation of component image velocity from local phase information. *Int. J. Comp. Vision*, 5:77–104, 1990.

[50] W. J. Hardcastle. *Physiology of Speech Production*. Academic Press, New York, NY, 1976.

[51] Hai Hong, Hartmut Neven, and Christoph von der Malsburg. Online facial expression recognition based on personalized galleries. In *Proc. of the 2nd Int. conf. on Automatic Face and Gesture Recognition*, pages 354–359, Nara, Japan, April 14-16 1998. IEEE.

[52] B. Horn and B. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[53] Wei Hua. Building facial expression analysis system. Technical report, CMU, Robotics Institute, 1998.

[54] T. Huang. *Face Recognition using 3-D Models*. NATO ASI Series F. Springer-Verlag, 1998.

[55] C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[56] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. *IEEE Trans. on Neural Networks*, 8(3):486–504, 1997.

[57] H. Kobayashi, K. Tange, and F. Hara. Real-time recognition of six basic facial expressions. In *Proceedings 4th IEEE International Workshop on Robot and Human Communication*, pages 179–186, Tokyo, Japan, 5-7 July 1995.

[58] Rob Koenen. Mpeg-4 overview - atlantic city version. In *ISO/IEC JTC1/SC29/WG11, Coding of Moving Pictures and Audio*, October 1998.

[59] A. Lanitis, C.J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 19, pages 743–756, 1997.

[60] James J. Lien, Takeo Kanade, Jeffrey F. Cohn, and Ching-Chung Li. Automated facial expression recognition based on facs action units. In *IEEE Proceedings of FG'98*, April 14-16 1998.

[61] James J. Lien, Takeo Kanade, Jeffrey F. Cohn, and Ching-Chung Li. Subtly different facial expression recognition and expression intensity estimation. In *IEEE Proceedings of CVPR'98, Santa Barbara, CA.*, June 1998.

[62] James J. Lien, Takeo Kanade, Adena J. Zlochower, Jeffrey F. Cohn, and Ching-Chung Li. Automatically recognizing facial expression in the spatio-temporal domain. In *Workshop on Perceptual User Interfaces, Banff, Alberta, Canada*, pages 94–97, October 19-21 1997.

[63] James Jenn-Jier Lien, Takeo Kanade, C. C. Li, and Jeffrey F. Cohn. A multi-method approach for discriminating between similar facial expressions, including expression intensity estimation. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 1998.

[64] Jenn-Jier James Lien. *Automatic Recognition of Facial Expression using Hidden Markov Models and Estimation of Expression Intensity.* PhD thesis, The Robotics Institute, CMU, April 1998.

[65] Christine L. Lisetti and David E. Rumelhart. Facial expression recognition using a neural network. In *Proceedings of the 11th International Flairs Conference*. AAAI Press, 1998.

[66] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *DARPA IU Workshop*, pages 121–130, 1981.

[67] Juergen Luettin, Neil A. Thacker, and Steve W. Beet. Active shape models for visual speech feature extraction. In D. G. Storck and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 383–390. Springer Verlag, Berlin, 1996.

[68] Juergen Luettin, Neil A. Thacker, and Steve W. Beet. Active shape models for visual speech feature extraction. Tech report No. 95/44, University of Sheffield, Electronic Systems Group, 1996.

[69] Juergen Luettin, Neil A. Thacker, and Steve W. Beet. Speachreading using shape and intensity information. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 58–61, 1996.

[70] Miachael Lyons and Shigeru Akamatsu. Coding facial expressions with gabor wavelets. In *IEEE Proceedings of the Second Int. Conf. on Automatic Face and Gesture Recognition, Nara, Japan*, pages 200–205, April 14-16 1998.

[71] K. Mase and A. Pentland. Recognition of facial expression from optical flow. In *IEICE Transactions*, volume E 74, pages 3474–3483, 1991.

[72] Baback Moghaddam, Wasiuddin Wahid, and Alex Pentland. Beyond eigenfaces: Probabilistic matching for face recognition. Tech Report 443, MIT Media Laboratory, Perceptual Computing Section, 1998.

[73] Yael Moses, David Reynard, and Andrew Blake. Robust real time tracking and classification of facial expressions. Technical report, Dept. of Engineering Science, University of Oxford, 1994.

[74] Shuichi Nishio, Kenji Koyama, and Toru Nakamura. Temporal differences in eye and mouth movements classifying facial expressions of smiles. In *IEEE Proceedings of the Sec. Int. Conf. on Automatic Face and Gesture Recognition, Nara, Japan*, pages 206–211, April 14-16 1998.

[75] Nuria Oliver, Alex Pentland, and Francois Berard. Lafter: Lips and face real time tracker with facial expression recognition. Tech Report No. 396, Mit Media Laboratory, Perceptual Computing Group, 1997.

[76] A. J. O'Toole, H. Abdi, K. A. Deffenbacher, and D. Valentin. A low-dimensional representaion of faces in the higher dimensions of the space. *J. Opt. Soc. Am.*, 10:405–411, 1993.

[77] Takahiro Otsuka and Jun Ohya. Spotting segments displaying facial expression from image sequences using hmm. In *IEEE Proceedings of the Sec. Int. Conf. on Automatic Face and Gesture Recognition, Nara, Japan*, pages 442–447, April 14-16 1998.

[78] C. Padgett and G. Cottrell. Representing face images for emotion classification. In M. Mozer, M. Jordan, and T. Petsche, editors, *Adcances in Neural Information Processing Systems*, volume 9, Cambridge, MA, 1997. MIT Press.

[79] Curtis Padgett and Garrison W. Cottrell. A simple neural network models categorical perception of facial expressions. In *Computer Science and Engineering, University of California, San Diego*, 1998.

[80] Curtis Padgett, Garrison W. Cottrell, and Ralph Adolphs. Categorical perception in facial emotion classification. Technical report, Computer Science and Engineering, University of California, San Diego, 1997.

[81] Barak A. Pearlmutter. Dynamic recurrent neural networks. Technical Report CMU-CS-90-196, CMU, December 1990.

[82] P. Penev and J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[83] J. Philips. Face recognition: Where are we now and where are we going. Panel Seession conducted at the International Conference on Automatic Face and Gesture Recognition, Killington, VT, October 1996.

[84] Frederic Pighin. Synthesizing realistic facial expressions from photographs. Technical report, University of Washington, 1998.

[85] S. M. Platt and N. I. Badler. Animating facial expressions. In *Computer Graphics*, volume 15, pages 245–252, August 1981.

[86] M. Rosenblum, Y. Yacoob, and L. Davis. Human expression recognition from motion using a radial basis function network architecture. In *IEEE Transactions on Neural Networks*, volume 7, pages 1121–1138, 1996.

[87] Mark Rosenblum, Larry Davis, and Yaser Yacoob. Human emotion recognition from motion using a radial basis function network architecture. In *IEEE Workshop on Motion of Non-rigid and Articulated Objects, Austin, TX*. Computer Vision Labratory, University of Maryland, November 1994.

[88] H.A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. Technical report CMU-CS-95-158r, Carnegie Mellon University, 1995.

[89] Henry A. Rowley, Baluja Shumeet, and Kanade Takeo. Rotation invariant neural network-based face detection. Technical Report CMU-CS-97-201, CS CMU, 1997.

[90] Ashok Samal and Prasana A. Iyengar. Automatic recognition and analysis of human faces and facial expression: A survey. In *Pattern Recognition*, volume 25, pages 65–77, 1992.

[91] L. C. De Silva, T. Miyasato, and R. Nakatsu. Facial emotion recognition using multimodal information. In *Int. Conf. on Information, Communications and Signal Processing (ICICS'97)*, pages 397–401, Singapore, September 1997. IEEE.

[92] E. P. Simoncelli. *Distributed Representation and Analysis of Visual Motion*. PhD thesis, Massachusetts Institute of Technology, 1993.

[93] A. Singh. *Optic Flow Computation*. IEEE Computer Society Press, Los Alamitos, CA, 1991.

[94] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4:519–524, 1987.

[95] M. Suwa, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. In *Proceedings of the 4th Int. Joint Conf. Pattern Recognition*, pages 408–410, 1978.

[96] Akikazu Takeuchi and Katashi Nagao. Communicative facial displays as a new conversational modality. In *INTERCHI '93*, 24-29 April 1993.

[97] D. Terzopoulos and K. Waters. Analysis of facial images using physical and anatomical models. In *Proceedings of the 3rd Int. Conf. on Computer vision*, pages 727–732, 1990.

[98] Nadia Magnenat Thalmann, Marc Escher, and Prem Kalra. Face to virtual face. Technical report, Miralab, CUI, University of Geneva, 1998.

[99] Dominique Valentin, Herve Abdi, Alice J. O'Toole, and Garrison W. Cottrell. Connectionist models of face processing: A survey. In *Pattern Recognition*, volume 27, pages 1209–1230, 1994.

[100] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 15, pages 733–741, 1997.

[101] Mei Wang, Yoshio Iwai, and Masahiko Yachida. Expression recognition from time-sequential facial images by use of expression change model. In *IEEE Proceedings of the Sec. Int. Conf. on Automatic Face and Gesture Recognition, Nara, Japan*, pages 324–329, April 14-16 1998.

[102] K. Waters. A muscle model for animating three-dimensional facial expression. In *Computer Graphics*, volume 22, pages 17–24, 1987.

[103] Y. T. Wu, T. Kanade, J. F. Cohn, and C. C. Li. Optical flow estimation using wavelet motion model. In *IEEE International Conference on Computer Vision*, pages 992–998, Bombay, India, January 1998.

[104] Yu-Te Wu, Takeo Kanade, Jeffrey Cohn, and Ching-Chung Li. Optical flow estimation using wavelet motion model. Technical report, The Robotics Institute, Carnegie Mellon University, 1997.

[105] Y. Yacoob and L. S. Davis. Recognizing human facial expression from long image sequences using optical flow. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 18, pages 636–642, 1996.

[106] Y. Yacoob, Heung-Man Lam, and Larry S. Davis. Recognizing faces showing expressions. Technical report, University of Maryland, Computer Vision Labratory,, 1995.

[107] Yaser Yacoob and Larry Davis. Computing spatio-temporal representations of human faces. Technical report, University of Maryland, Computer Vision Lab, 1994.

[108] Yaser Yacoob and Larry Davis. Recognizing facial expression by spatio-temporal analysis. Technical report, University of Maryland, Computer Vision Labratory, 1994.

[109] Zhengyou Zhang, Michael Lyons, Michael Schuster, and Shigeru Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *IEEE Proceedings of the Second Int. Conf. on Automatic Face and Gesture Recognition (FG'98), Nara, Japan*, pages 454–459, April 14-16 1998.

## 9.2 Psychology

[110] E. Baenninger-Huber. Prototypical affective microsequences in psychotherapeutic interaction. In P. Ekman and E. L. Rosenberg, editors, *What the face reveals*, Oxford, 1998. Oxford University Press.

[111] Eva Baenninger-Huber. Prototypical affective microsequences in psychotherapeutic interaction. In *Psychotherapy Research*, volume 2, pages 291–306, 1992.

[112] Eva Baenninger-Huber and Felix Steiner. Facs in psychotherapy reserach. Technical Report 19, Abteilung Klinische Psychologie, Universitaet Zuerich, Schmelzbergstrasse 40 8044 Zuerich, Switzerland, 1986.

[113] Jeffrey F. Cohn and Gary S. Katz. Bimodal expression of emotion by face and voice. In *Department of Psychology, University of Pittsburgh and Robotics Institute, CMU*, 1998.

[114] Jeffrey F. Cohn, Adena J. Zlochower, James Lien, and Takeo Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. In *Department of Psychology, University of Pittsburgh and Robotics Institute, CMU*, 1998.

[115] Charles Darwin. *The Expression of the Emotions in Man and Animals*. J. Murray, London, 1872.

[116] P. Ekman. About brows: Emotional and conversational signals. In *Human Ethology*, pages 169–202. J. Aschoff, M. Con Carnach, K. Foppa, W. Lepenies and D. Plog, 1979.

[117] P. Ekman. Emotions in the human face. *Cambridge University Press*, 1982.

[118] P. Ekman. Expression and the nature of emotion. *Lawrence Erlbaum*, pages 319–343, 1984.

[119] P. Ekman. The argument and evidence about universals in facial expressions of emotion. In D. Raskin, editor, *Psychological Methods in Criminal Investigation and Evidence*, pages 297–332. Springer Publishing Co, inc., New York, 1989.

[120] P. Ekman. Facial expressions of emotions: New findings, new questions. In *Psychological Science*, volume 3, pages 34–38, 1992.

[121] P. Ekman, W. Friesen, and M. O'Sullivan. Smiles when lying. *Journal of Personality and Social Psychology*, 545:414–420, 1988.

[122] P. Ekman and W. V. Friesen. Manual for the facial action coding system. *Consulting Psychologists Press, Palo Alto*, 1977.

[123] P. Ekman and W. V. Friesen. Manual for the facial action coding system. *Palo Alto: Consulting Psychologist Press*, 1978.

[124] P. Ekman, Thomas S. Huang, Terrence J. Sejnowski, and J. C. Hager. Final report to nsf of the planning workshop on facial expression. Ftp helmholtz.sdsc.edu, 1992.

[125] W. V. Friesen and P. Ekman. Emotional facial action coding system - 7. Unpublished Manual, 1984.

[126] J. Hager and P. Ekman. The essential behavioral science of the face and gesture that computer scientists need to know. In M. Bichsel, editor, *Proceedings of the International Workshop on Automatic Face- and Gesture Recognition*, pages 7–11, Winterthurerstrasse 190, CH-8057, Switzerland, 1995. University of Zurich, Department of Computer Science.

[127] C. Izard. The maximally descriminative facial movement coding system (max). Available from Instructional Resource Center, University of Delaware, Newark, Delaware, 1979.

[128] S. Kaiser and T. Wehrle. Automated coding of facial behavior in human-computer interactions with facs. In *Journal of Nonverbal Behavior*, volume 16, pages 67–83, 1992.

[129] S. Kaiser and T. Wehrle. Situated emotional problem solving in interactive computer games. In *Proceedings of the VIXth Conference of the International Society for Reserach on Emotions, ISRE Publications*, 1996.

[130] S. Kaiser, T. Wehrle, and P. Edwards. Multi-modal emotion measurement in an interactive computer-game: A pilot-study. In *Proceedings of the VIIIth Conference of the International Society for Research on Emotions, ISRE Publications*, 1994.

[131] S. Kaiser, T. Wehrle, and S. Schmidt. Emotional episodes, facial expressions, and reported feelings in human-computer interactions. In *Proceedings of the Xth conference of the international society for research on emotions, isre publications*, 1998.

[132] Susanne Kaiser and Thomas Wehrle. Emotion research and ai: Some theoretical and technical issues. http://www.unige.ch/fapse/emotion/members/kaiser/rai4.htm, 1994.

[133] J. L. McClelland and D. E. Rumelhart. A distributed model of human learning and memory. In J. L. McClelland and D. E. Rumelhart, editors, *Parallel Distributed Processing. Exploration in the Microstructure of Cognition*, volume 2, 1986.

[134] K. R. Scherer. On the nature and function of emotion: A component process approach. In K. R. Scherer and P. Ekamn, editors, *Approaches to Emotion, Lawrence Erlbaum*, pages 293–318, 1984.

[135] H. Schlosberg. Three dimensions of emotion. In *Psychol. Rev.*, volume 61, pages 81–88, 1954.

[136] Maria Von Salisch and Eva Baenninger-Huber. Investigating the facial expression of emotion in face-to-face interaction. In *Proceedings of the VIII conference of the International Society for Research on Emotions, ISRE Publications*, 1994.

[137] T. Wehrle, S. Schmidt, S. Kaiser, and K. R. Scherer. Studying dynamic models of facial expression of emotion using synthetic animated faces. In *Journal of Personality and Social Psychology*, 1998.

[138] Thomas Wehrle. Motivations behind modeling emotional agents: Whose emotion does your robot have. In *5th International Conference of the Society for Adaptive Behavior, SAB98*, 1998.

## 9.3 Neuroscience


[139] V. Bruce. Recognizing faces. In *Lawrence Erlbaum Association, London*, 1988.

[140] V. Bruce. Stability from variation: The case of face recognition. *Quarterly Journal of Experimental Psychology*, 47A:5–28, 1994.

[141] V. Bruce. Human face perception and identification. In H. Wechsler, P. Phillips, V. Bruce, F. Fogelman-Soulie, and T. Huang, editors, *Face Recognition: From Theory to Applications*, Nato ASI Series F. Springer-Verlag, 1998. In press.

[142] R. Cabeza, V. Bruce, T. Kato, and M. Oda. The prototype effect in face recognition: Extension and limits. *Memory and Cognition*, 1998.

[143] P. S. Churchland and T. J. Sejnowski. *The Computational Brain*. MIT Press, Cambridge: Ma, 1992.

[144] H. D. Ellis. Processes underlying face recognition. In R. Bruyer, editor, *The Neuropsychology of Face Perception and Facial Expression*, volume Chapter 1, pages 1–27. Lawrence Erlbaum Associates, New Jersey, 1986.

[145] Geoffrey E. Hinton. Wie neuronale netze aus erfahrung lernen. In *Spektrum der Wissenschaft, Gehirn und Bewusstsein*, pages 136–145, November 1992.

[146] M. H. Johnson and J. Morton. Biology and cognitive development: The case of face recognition. Oxford, UK; Cambridge, Mass: Blackwell, 1991.

[147] D. Pollen and S. Ronner. Phase relationship between adjacent simple cells in the visula cortex. *Science*, 212:1409–1411, 1981.

[148] M. P. Young and S. Yamane. Sparse population coding of faces in the inferotemporal cortex. *Science*, 56:1327–1331, 1992.


## 9.4 Medicine


[149] R. Buck and R. Duffy. Nonverbal communication of affect in brain-damaged patients. *Cortex*, 16:351–362, 1980.

[150] Pavel Dulguerov. Review of objective topographic facial nerve evalutation methods. Technical report, Department of Otolaryngology -Geneva University Hospital, 24, rue Micheli-du-Crest 1211 Geneva 14, Switzerland, 1998.

[151] J.W. House. Facial nerve grading systems. *Laryngoscope*, 93:1056–1069, 1983.

[152] J.W. House and D. E. Brackmann. Facial nerve grading system. *Otolaryngol Head Neck Surg*, 93:146–147, 1985.

[153] R. Krause, E. Steimer, C. Sanger-Alt, and G. Wagner. Facial expression and schizophrenic patients and their interaction partners. *Psychiatry*, 52:1–12, 1989.

[154] B. G. Ross, G. Fradet, and J. M. Nedzelski. Development of a sensitive clinical grading system. *Otolaryngol Head Neck Surg*, 114:380–386, 1996.

[155] E. Steimer-Krause, R. Krause, and G. Wagner. Interaction regualtions used by schizophrenics and psychosomatic patients. studies on facial behavior in dyadic interactions. *Psychiatry*, 53:209–228, 1990.