Automatic Analysis of Facial Expressions: The State of the Art

Maja Pantic, *Student Member*, *IEEE*, and Leon J.M. Rothkrantz

Abstract—Humans detect and interpret faces and facial expressions in a scene with little or no effort. Still, development of an automated system that accomplishes this task is rather difficult. There are several related problems: detection of an image segment as a face, extraction of the facial expression information, and classification of the expression (e.g., in emotion categories). A system that performs these operations accurately and in real time would form a big step in achieving a human-like interaction between man and machine. This paper surveys the past work in solving these problems. The capability of the human visual system with respect to these problems is discussed, too. It is meant to serve as an ultimate goal and a guide for determining recommendations for development of an automatic facial expression analyzer.

Index Terms—Face detection, facial expression information extraction, facial action encoding, facial expression emotional classification.

1 Introduction

As pointed out by Bruce [6], Takeuchi and Nagao [84], and Hara and Kobayashi [28], human face-to-face communication is an ideal model for designing a multimodal/media human-computer interface (HCI). The main characteristics of human communication are: multiplicity and multimodality of communication channels. A channel is a communication medium while a modality is a sense used to perceive signals from the outside world. Examples of human communication channels are: auditory channel that carries speech, auditory channel that carries vocal intonation, visual channel that carries facial expressions, and visual channel that carries body movements. The senses of sight, hearing, and touch are examples of modalities. In usual face-to-face communication, many channels are used and different modalities are activated. As a result, communication becomes highly flexible and robust. Failure of one channel is recovered by another channel and a message in one channel can be explained by another channel. This is how a multimedia/modal HCI should be developed for facilitating robust, natural, efficient, and effective man-machine interaction.

Relatively few existing works combine different modalities into a single system for human communicative reaction analysis. Examples are the works of Chen et al. [9] and De Silva et al. [15] who studied the effects of a combined detection of facial and vocal expressions of emotions. So far, the majority of studies treat various human communication channels separately, as indicated by Nakatsu [58]. Examples for the presented systems are: emotional interpretation of human voices [35], [66], [68], [90], emotion recognition by physiological signals pattern recognition [67], detection and interpretation of hand gestures [64], recognition of body movements [29], [97], [46], and facial expression analysis (this survey).

The terms "face-to-face" and "interface" indicate that the face plays an essential role in interpersonal communication. The face is the mean to identify other members of the species, to interpret what has been said by the means of lipreading, and to understand someone's emotional state and intentions on the basis of the shown facial expression. Personality, attractiveness, age, and gender can also be seen from someone's face. Considerable research in social psychology has also shown that facial expressions help coordinate conversation [4], [82], and have considerably more effect on whether a listener feels liked or disliked than the speaker's spoken words [55]. Mehrabian indicated that the verbal part (i.e., spoken words) of a message contributes only for 7 percent to the effect of the message as a whole, the vocal part (e.g., voice intonation) contributes for 38 percent, while facial expression of the speaker contributes for 55 percent to the effect of the spoken message [55]. This implies that the facial expressions form the major modality in human communication.

Recent advances in image analysis and pattern recognition open up the possibility of automatic detection and classification of emotional and conversational facial signals. Automating facial expression analysis could bring facial expressions into man-machine interaction as a new modality and make the interaction tighter and more efficient. Such a system could also make classification of facial expressions widely accessible as a tool for research in behavioral science and medicine. The goal of this paper is to survey the work done in automating facial expression analysis in facial images and image sequences. Section 2 identifies three basic problems related to facial expression analysis. These problems are: face detection in a facial image or image sequence, facial expression data extraction, and facial expression classification. The capability of the

• *The authors are with the Department of Media, Engineering and Mathematics, Delft University of Technology, PO Box 356, 2600 AJ Delft, The Netherlands.*
E-mail: {M.Pantic, L.J.M.Rothkrantz}@cs.tudelft.nl.

Manuscript received 29 June 1999; revised 16 March 2000; accepted 12 Sept. 2000.
Recommended for acceptance by K. Bowyer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 110156.

human visual system with respect to these problems is described. It defines, in some way, the expectations for an automated system. The characteristics of an ideal automated system for facial expression analysis are given in Section 3. Section 4 surveys the techniques presented in the literature in the past decade for facial expression analysis by a computer. Their characteristics are summarized in respect to the requirements posed on the design of an ideal facial expression analyzer. We do not attempt to provide an exhaustive review of the past work in each of the problems related to automatic facial expression analysis. Here, we selectively discuss systems which deal with each of these problems. Possible directions for future research are discussed in Section 5. Section 6 concludes the paper.

2 Facial Expression Analysis

Our aim is to explore the issues in design and implementation of a system that could perform automated facial expression analysis. In general, three main steps can be distinguished in tackling the problem. First, before a facial expression can be analyzed, the face must be detected in a scene. Next is to devise mechanisms for extracting the facial expression information from the observed facial image or image sequence. In the case of static images, the process of extracting the facial expression information is referred to as *localizing* the face and its features in the scene. In the case of facial image sequences, this process is referred to as *tracking* the face and its features in the scene. At this point, a clear distinction should be made between two terms, namely, facial features and face model features. The *facial features* are the prominent features of the face—eyebrows, eyes, nose, mouth, and chin. The *face model features* are the features used to represent (model) the face. The face can be represented in various ways, e.g., as a whole unit (*holistic* representation), as a set of features (*analytic* representation) or as a combination of these (*hybrid* approach). The applied face representation and the kind of input images determine the choice of mechanisms for automatic extraction of facial expression information. The final step is to define some set of categories, which we want to use for facial expression classification and/or facial expression interpretation, and to devise the mechanism of categorization.

Before an automated facial expression analyzer is built, one should decide on the system's functionality. A good reference point is the functionality of the human visual system. After all, it is the best known facial expression analyzer. This section discusses the three basic problems related to the process of facial expression analysis as well as the capability of the human visual system with respect to these.

2.1 Face Detection

For most works in automatic facial expression analysis, the conditions under which a facial image or image sequence is obtained are controlled. Usually, the image has the face in frontal view. Hence, the presence of a face in the scene is ensured and some global location of the face in the scene is known a priori. However, determining the exact location of the face in a digitized facial image is a more complex problem. First, the scale and the orientation of the face can vary from image to image. If the mugshots are taken with a fixed camera, faces can occur in images at various sizes and angles due to the movements of the observed person. Thus, it is difficult to search for a fixed pattern (template) in the image. The presence of noise and occlusion makes the problem even more difficult.

Humans detect a facial pattern by casual inspection of the scene. We detect faces effortlessly in a wide range of conditions, under bad lightning conditions or from a great distance. It is generally believed that two-gray-levels images of 100 to 200 pixels form a lower limit for detection of a face by a human observer [75], [8]. Another characteristic of the human visual system is that a face is perceived as a whole, not as a collection of the facial features. The presence of the features and their geometrical relationship with each other appears to be more important than the details of the features [5]. When a face is partially occluded (e.g., by a hand), we perceive a whole face, as if our perceptual system fills in the missing parts. This is very difficult (if possible at all) to achieve by a computer.

2.2 Facial Expression Data Extraction

After the presence of a face has been detected in the observed scene, the next step is to extract the information about the encountered facial expression in an automatic way. If the extraction cannot be performed automatically, a fully automatic facial expression analyzer cannot be developed. Both, the applied face representation and the kind of input images affect the choice of the approach to facial expression information extraction.

One of the fundamental issues about the facial expression analysis is the representation of the visual information that an examined face might reveal [102]. The results of Johansson's point-light display experiments [1], [5], gave a clue to this problem. The experiments suggest that the visual properties of the face, regarding the information about the shown facial expression, could be made clear by describing the movements of points belonging to the facial features (eyebrows, eyes, and mouth) and then by analyzing the relationships between those movements. This triggered the researchers of vision-based facial gesture analysis to make different attempts to define point-based visual properties of facial expressions. Various analytic face representations yielded, in which the face is modeled as a set of facial points (e.g., Fig. 6, [42], Fig. 7, [61]) or as a set of templates fitted to the facial features such as the eyes and the mouth. In another approach to face representation (holistic approach), the face is represented as a whole unit. A 3D wire-frame with a mapped texture (e.g., [86]) and a spatio-temporal model of facial image motion (e.g., Fig. 8, [2]) are typical examples of the holistic approaches to face representation. The face can be also modeled using a so-called hybrid approach, which typifies a combination of analytic and holistic approaches to face representation. In this approach, a set of facial points is usually used to determine an initial position of a template that models the face (e.g., Fig 10, [40]).

Irrespectively of the kind of the face model applied, attempts must be made to model and then extract the information about the displayed facial expression without losing any (or much) of that information. Several factors make this task complex. The first is the presence of facial

hair, glasses, etc., which obscure the facial features. Another problem is the variation in size and orientation of the face in input images. This disables a search for fixed patterns in the images. Finally, noise and occlusion are always present to some extent.

As indicated by Ellis [23], human encoding of the visual stimulus (face and its expression) may be in the form of a primal sketch and may be hardwired. However, not much else is known in terms of the nature of internal representation of a face in the human brain.

2.3 Facial Expression Classification

After the face and its appearance have been perceived, the next step of an automated expression analyzer is to "identify" the facial expression conveyed by the face. A fundamental issue about the facial expression classification is to define a set of categories we want to deal with. A related issue is to devise mechanisms of categorization. Facial expressions can be classified in various ways—in terms of facial actions that cause an expression, in terms of some nonprototypic expressions such as "raised brows" or in terms of some prototypic expressions such as emotional expressions.

The *Facial Action Coding System* (FACS) [21] is probably the most known study on facial activity. It is a system that has been developed to facilitate objective measurement of facial activity for behavioral science investigations of the face. FACS is designed for human observers to detect independent subtle changes in facial appearance caused by contractions of the facial muscles. In a form of rules, FACS provides a linguistic description of all possible, visually detectable, facial changes in terms of 44 so-called *Action Units* (AUs). Using these rules, a trained human FACS coder decomposes a shown expression into the specific AUs that describe the expression. Automating FACS would make it widely accessible as a research tool in the behavioral science, which is furthermore the theoretical basis of multimodal/media user interfaces. This triggered researchers of computer vision field to take different approaches in tackling the problem. Still, explicit attempts to automate the facial action coding as applied to automated FACS encoding are few (see [16] or [17] for a review of the existing methods as well as Table 7 of this survey).

Most of the studies on automated expression analysis perform an emotional classification. As indicated by Fridlund et al. [25], the most known and the most commonly used study on emotional classification of facial expressions is the cross-cultural study on existence of "universal categories of emotional expressions." Ekman defined six such categories, referred to as the *basic emotions*: happiness, sadness, surprise, fear, anger, and disgust [19]. He described each basic emotion in terms of a facial expression that uniquely characterizes that emotion. In the past years, many questions arose around this study. Are the basic emotional expressions indeed universal [33], [22], or are they merely a stressing of the verbal communication and have no relation with an actual emotional state [76], [77], [26]? Also, it is not at all certain that each facial expression able to be displayed on the face can be classified under the six basic emotion categories. Nevertheless, most

Fig. 1. Expressions of blended emotions (surprise—happiness).

of the studies on vision-based facial expression analysis rely on Ekman's emotional categorization of facial expressions.

The problem of automating facial expression emotional classification is difficult to handle for a number of reasons. First, Ekman's description of the six prototypic facial expressions of emotion is linguistic (and, thus, ambiguous). There is no uniquely defined description either in terms of facial actions or in terms of some other universally defined facial codes. Hence, the validation and the verification of the classification scheme to be used are difficult and crucial tasks. Second, classification of facial expressions in to multiple emotion categories should be feasible (e.g., raised eyebrows and smiling mouth is a blend of surprise and happiness, Fig. 1). Still, there is no psychological scrutiny on this topic.

Three more issues are related to facial expression classification in general. First, the system should be capable of analyzing any subject, male or female of any age and ethnicity. In other words, the classification mechanism may not depend on physiognomic variability of the observed person. On the other hand, each person has his/her own maximal intensity of displaying a particular facial expression. Therefore, if the obtained classification is to be quantified (e.g., to achieve a quantified encoding of facial actions or a quantified emotional labeling of blended expressions), systems which can start with a generic expression classification and then adapt to a particular individual have an advantage. Second, it is important to realize that the interpretation of the body language is situation-dependent [75]. Nevertheless, the information about the context in which a facial expression appears is very difficult to obtain in an automatic way. This issue has not been handled by the currently existing systems. Finally, there is now a growing psychological research that argues that timing of facial expressions is a critical factor in the interpretation of expressions [1], [5], [34]. For the researchers of automated vision-based expression analysis, this suggests moving towards a real-time whole-face analysis of facial expression dynamics.

While the human mechanisms for face detection are very robust, the same is not the case for interpretation of facial expressions. It is often very difficult to determine the exact nature of the expression on a person's face. According to Bassili [1], a trained observer can correctly classify faces

showing six basic emotions with an average of 87 percent. This ratio varies depending on several factors: the familiarity with the face, the familiarity with the personality of the observed person, the general experience with different types of expressions, the attention given to the face and the nonvisual cues (e.g., the context in which an expression appears). It is interesting to note that the appearance of the upper face features plays a more important role in face interpretation as opposed to lower face features [22].

3 An Ideal System for Facial Expression Analysis

Before developing an automated system for facial expression analysis, one should decide on its functionality. A good reference point is the best known facial expression analyzer—the human visual system. It may not be possible to incorporate all features of the human visual system into an automated system, and some features may even be undesirable, but it can certainly serve as a reference point.

A first requirement that should be posed on developing an ideal automated facial expression analyzer is that all of the stages of the facial expression analysis are to be performed automatically, namely, face detection, facial expression information extraction, and facial expression classification. Yet, actual implementation and integration of these stages into a system are constrained by the system's application domain. For instance, if the system is to be used as a tool for research in behavioral science, real-time performance is not an essential property of the system. On the other hand, this is crucial if the system would form a part of an advanced user-interface. Long delays make the interaction desynchronized and less efficient. Also, having an explanation facility that would elucidate facial action encoding performed by the system might be useful if the system is employed to train human experts in using FACS. However, such facility is superfluous if the system is to be employed in videoconferencing or as a stress-monitoring tool. In this paper, we are mainly concerned with two major application domains of an automated facial expression analyzer, namely, behavioral science research and multimodal/media HCI. In this section, we propose an ideal automated facial expression analyzer (Table 1) which could be employed in those application domains and has the properties of the human visual system.

Considering the potential applications of an automated facial expression analyzer, which involve continuous observation of a subject in a time interval, facial image acquisition should proceed in an automatic way. In order to be universal, the system should be capable of analyzing subjects of both sexes, of any age and any ethnicity. Also, no constraints should be set on the outlook of the observed subjects. The system should perform robustly despite changes in lightning conditions and distractions like glasses, changes in hair style, and facial hair like moustache, beard and grown-together eyebrows. Similarly to the human visual system, an ideal system would "fill in" missing parts of the observed face and "perceive" a whole face even when a part of it is occluded (e.g., by hand). In most real-life situations, complete immovability of the observed subject cannot be assumed. Hence, the system should be able to deal with rigid head motions. Ideally, the system would perform robust facial expression analysis despite large changes in viewing conditions; it would be capable of dealing with a whole range of head movements, from frontal view to profile view acquired by a fixed camera. This could be achieved by employing several fixed cameras for acquiring different facial views of the examined face (such as frontal view, and right and left profile views) and then approximating the actual view by interpolation among the acquired views. Having no constraints set on the rigid head motions of the subject can also be achieved by having a camera mounted on the subject's head and placed in front of his/her face.

TABLE 1
Properties of an Ideal Analyzer

	General Characteristic
1	Automatic facial image acquisition
2	Subjects of any age, ethnicity and outlook
3	Deals with variation in lightning
4	Deals with partially occluded faces
5	No special markers/make-up required
6	Deals with rigid head motions
7	Automatic face detection
8	Automatic facial expression data extraction
9	Deals with inaccurate facial expression data
10	Automatic facial expression classification
11	Distinguishes all possible expressions
12	Deals with unilateral facial changes
13	Obeys anatomical rules (see [21])
	Behavioral science research application
14	Distinguishes all 44 facial actions ([21])
15	Quantifies facial action codes
	Multi-modal/media HCI application
16	# interpretation categories unlimited
17	Features adaptive learning facility
18	Assigns quantified interpretation labels
19	Assigns multiple interpretation labels
20	Features real-time processing

An ideal system should perform robust automatic face detection and facial expression information extraction in the acquired images or image sequences. Considering the state-of-the-art in image processing techniques, inaccurate, noisy, and missing data could be expected. An ideal system should be capable of dealing with these inaccuracies. In addition, certainty of the extracted facial expression information should be taken into account.

An ideal system should be able to perform analysis of all visually distinguishable facial expressions. Well-defined face representation is a prerequisite for achieving this. The face representation should be such that a particular alteration of the face model uniquely reveals a particular facial expression. In general, an ideal system should be able to distinguish:

1. all possible facial expressions (a reference point is a total of 44 facial actions defined in FACS [21]

TABLE 2
Early Methods for Automatic Facial Expression Analysis

Reference	Characteristics of an ideal automated facial expression analyzer (Table 1)																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	17	18	19
Analysis from static facial images																		
Cottrell [13]	×	●	×	×	●	–	×	–	–	●	×	●	–	×	8	×	×	×
Kearney [40]	×	●	×	×	●	–	×	×	–	●	×	●	●	36	n	●	×	●
Hara [43]	×	●	×	×	●	–	×	×	–	●	×	●	●	×	6	×	●	●
Matsuno [54]	×	×	●	×	●	–	×	×	–	●	×	●	●	×	4	×	×	×
Rahardja [70]	×	–	×	–	–	–	×	–	–	●	×	–	–	×	6	×	×	×
Ushida [92]	×	●	×	×	●	–	×	×	–	●	×	×	●	×	3	×	×	×
Vanger [93]	×	●	×	×	●	–	×	×	–	●	×	●	●	×	7	×	×	×
Analysis from facial image sequences																		
Mase [52]	×	–	×	×	●	×	×	●	×	●	×	●	●	×	4	×	×	×
Moses [57]	●	–	●	×	●	●	●	●	●	●	×	●	●	5	5	×	×	×
Rosenblum [73]	×	–	–	×	●	●	×	●	●	●	×	●	●	×	2	×	×	×
Yacoob [100]	–	–	–	×	●	●	×	●	●	●	×	●	●	×	7	×	●	×

Legend: ● = "yes", × = "no", – = missing entry

whose combinations form the complete set of facial expressions),
2. any bilateral or unilateral facial change [21], and
3. facial expressions with a similar facial appearance (e.g., upward pull of the upper lip and nose wrinkling which also causes the upward pull of the upper lip [21]).

In practice, it may not be possible to define a face model that can satisfy both, to reflect each and every change in facial appearance and whose features are detectable in a facial image or image sequence. Still, the set of distinct facial expressions that the system can distinguish should be as copious as possible.

If the system is to be used for behavioral science research purposes it should perform facial expression recognition as applied to automated FACS encoding. As explained by Bartlett et al. [16], [17], this means that it should accomplish multiple quantified expression classification in terms of 44 AUs defined in FACS. If the system is to be used as a part of an advanced multimodal/media HCI, the system should be able to interpret the shown facial expressions (e.g., in terms of emotions). Since psychological researchers disagree on existence of universal categories of emotional facial displays, an ideal system should be able to adapt the classification mechanism according to the user's subjective interpretation of expressions, e.g., as suggested in [40]. Also, it is definitely not the case that each and every facial expression able to be displayed on the face can be classified into one and only one emotion class. Think about blended emotional displays such as raised eyebrow and smiling mouth (Fig. 1). This expression might be classified in two emotion categories defined by Ekman and Friesen [20]—surprise and happiness. Yet, according to the descriptions of these prototypic expressions given by Ekman and Friesen [20], the left hand side facial expression shown in Fig. 1 belongs "more" to the surprise—than to the happiness class. For instance, in the left hand side image the "percentage" of shown surprise is higher than the "percentage" of shown happiness while those percentages are approximately the same in the case of the right hand side image. In order to obtain an accurate categorization, an ideal analyzer should perform quantified classification of facial expression into multiple emotion categories.

4 Automatic Facial Expression Analysis

For its utility in application domains of human behavior interpretation and multimodal/media HCI, automatic facial expression analysis has attracted the interest of many computer vision researchers. Since the mid 1970s, different approaches are proposed for facial expression analysis from either static facial images or image sequences. In 1992, Samal and Iyengar [79] gave an overview of the early works. This paper explores and compares approaches to automatic facial expression analysis that have been developed recently, i.e., in the late 1990s. Before surveying these works in detail, we are giving a short overview of the systems for facial expression analysis proposed in the period of 1991 to 1995.

Table 2 summarizes the features of these systems in respect to the requirements posed on design of an ideal facial expression analyzer. None of these systems performs a quantified expression classification in terms of facial actions. Also, except the system proposed by Moses et al. [57], no system listed in Table 2 performs in real-time. Therefore, these properties (i.e., columns 15 and 20) have been excluded from Table 2 (● stands for "yes," **X** stands for "no," and - represents a missing entry). A missing entry either means that it is not reported on the issue or that the issue is not applicable to the system in question. An inapplicable issue, for instance, is the issue of dealing with rigid head motions and inaccurate facial data in the cases where the input data were hand measured (e.g., [40]). Some of the methods listed in Table 2 do not perform automatic facial data extraction (see column 8); the others achieve this by using facial motion analysis [52], [100], [101], [73], [57]. Except the method proposed by Kearney and McKenzie [40], which performs facial expression classification in

TABLE 3
Recent Approaches to Automatic Facial Expression Analysis

Reference	Characteristics of an ideal automated facial expression analyzer (Table1)																	
	1	2	3	5	6	7	8	9	10	11	12	13	14	16	17	18	19	20
Analysis from static facial images																		
Edwards [18]	●	-	●	●	●	×	●	-	●	×	●	●	×	7	×	×	×	×
Hara [42]	●	1	-	●	×	●	●	-	●	×	●	-	×	6	×	×	×	●
Hong [30]	●	-	×	●	●	●	●	●	●	×	●	●	×	7	×	×	×	●
Huang [32]	●	1	×	●	×	●	●	-	●	×	●	●	×	6	×	×	×	×
Lyons [51]	×	●	×	●	-	×	×	-	●	×	●	●	×	7	×	×	×	×
Padget [61]	×	●	×	●	-	×	×	-	●	×	●	●	×	7	×	×	×	×
Pantic [62]	●	3	×	●	●	●	●	●	●	×	●	●	31	6	×	●	●	×
Yoneyama [104]	●	1	-	●	×	●	●	-	●	×	●	●	×	4	×	×	×	-
Zhang [106]	×	●	×	●	-	×	×	-	●	×	●	●	×	7	×	●	●	×
Zhao [107]	×	●	×	●	-	×	×	-	●	×	●	-	×	6	×	×	×	×
Analysis from facial image sequences																		
Black [2]	●	-	●	●	●	×	●	×	●	×	●	●	-	6	×	●	×	×
Cohn [10]	●	3	×	●	×	×	●	×	●	×	●	●	15	-	×	×	×	-
Essa [24]	●	●	●	●	-	●	●	●	●	×	●	●	2	4	×	×	×	●
Kimura [41]	●	×	●	●	×	●	●	●	●	×	●	●	×	3	×	●	×	-
Otsuka [60]	●	-	-	●	●	-	●	×	●	×	×	●	×	6	×	×	×	×
Wang [95]	●	1	×	●	×	×	●	-	●	×	●	●	×	3	×	●	×	×

Legend: ● = "yes", × = "no", - = missing entry

terms of facial actions as well as in terms of interpretation categories defined by users, the systems listed in Table 2 perform expression classification into a number of basic emotion categories. The utilized techniques include holistic spatial analysis [13], [70], [54], spatio-temporal analysis [52], [100], [73], [57], and analytic spatial analysis [93], [43], [44], [92], [40]. The correct recognition rates reported range from 70-92 percent when recognizing 3-8 emotion categories.

The survey of the works presented in the literature between 1996 and 2000 is divided into three parts, based on the problems discussed in Section 2—face detection, facial expression information extraction, and facial expression classification. We do not attempt to provide an exhaustive review of the past work in each of the problems related to automatic facial expression analysis. Here, we selectively discuss recently developed systems, which deal with both, facial expression detection and classification. Table 3 summarizes the properties of the surveyed facial expression analyzers in respect to the requirements posed on design of an ideal facial expression analyzer (Table 1). None of these systems are able to perform facial expression information extraction from images of occluded faces or a quantified expression classification in terms of facial actions. Therefore, these properties (i.e., columns 4 and 15) have been excluded from Table 3. In the case of systems where facial expression information is manually extracted, we are declaring that those can deal with the subjects of any ethnicity. In the case of automatic facial expression data extraction, the value of the column 2 in Table 3 represents the range in ethnicity of testing subjects. The number of testing images, the number of subjects used to make the testing images, and the overall performance of the surveyed systems is summarized in Table 8.

The approaches that have been explored lately also include systems for automatic analysis and synthesis of facial expressions [86], [56], [39], [47], [89], [88], [53], [14], [19] (for a broader list of references see [87]). Although the image analysis techniques in these systems are relevant to the present goals, the systems themselves are of limited use for behavioral science investigations of the face and for multimodal/media HCI. These systems primarily concern facial expression animation and do not attempt to classify the observed facial expression either in terms of facial actions or in terms of emotion categories. For this reason, these and similar methods are out of the scope of this paper, which goal is to explore and compare image-based approaches to facial expression detection and classification.

4.1 Face Detection

For most of the work in automatic facial expression analysis, the conditions under which an image is obtained are controlled. The camera is either mounted on a helmet-like device worn by the subject (e.g., [62], [59]) or placed in such a way that the image has the face in frontal view. Hence, the presence of the face in the scene is ensured and some global location of the face in the scene is known a priori. Yet, in most of the real-life situations where an automated facial expression analyzer is to be employed (e.g., in a multimodal/media HCI), the location of a face in the image is not known a priori. Recently, the problem of automatic face detection in an arbitrary scene has drawn great attention (e.g., see [74], [83], [85]).

Independently of the kind of input images—facial images or arbitrary images—detection of the exact face position in an observed image or image sequence has been approached in two ways. In the holistic approach, the face is determined as a whole unit. In the second, analytic

TABLE 4
Summary of the Methods for Automatic Face Detection

	Reference	View	Method	Comments
Facial images				
Holistic approach	Huang [32]	Frontal view	Canny edge detector PDM model fitting	No rigid head rotations
	Pantic [62]	Dual view	Image histogram analysis Thresholding	Mounted camera on the subject's head
Analytic approach	Hara [42]	Frontal view	Brightness distribution	No rigid head motions Real-time process
	Yoneyama [104]	Frontal view	-	-
	Kimura [41]	Frontal view	Integral projection [99] Potential Net fitting	No rigid head rotation
Arbitrary images				
Holistic approach	Hong [30]	Frontal view	Steffens et al. [81]: Spatio-temporal filtering Stereo algorithm Color detector Convex region detector Linear predictive filter	Complex background Slight head motions
	Essa [24]	Frontal to profile view	Pentland et al. [65]: Spatiotemporal filtering Eigenfaces Eigenfeatures	Complex background Rigid head motions Faces with facial hair Faces with glasses Real-time process

approach, the face is detected by detecting some important facial features first (e.g., the irises and the nostrils). The location of the features in correspondence with each other determines then the overall location of the face. Table 4 provides a classification of facial expression analyzers according to the kind of input images and the applied method.

4.1.1 Face Detection in Facial Images

To represent the face, Huang and Huang [32] apply a point distribution model (PDM). In order to achieve a correct placement of an initial PDM in an input image, Huang and Huang utilize a Canny edge detector to obtain a rough estimate of the face location in the image. The valley in pixel intensity that lies between the lips and the two symmetrical vertical edges representing the outer vertical boundaries of the face generate a rough estimate of the face location. The face should be without facial hair and glasses, no rigid head motion may be encountered and illumination variations must be linear for the system to work correctly.

Pantic and Rothkrantz [62] detect the face as a whole unit, too. As input to their system, they use dual-view facial images. To determine the vertical and horizontal outer boundaries of the head, they analyze the vertical and horizontal histogram of the frontal-view image. To localize the contour of the face, they use an algorithm based on the HSV color model, which is similar to the algorithm based on the relative RGB model [103]. For the profile view image they apply a profile-detection algorithm, which represents a spatial approach to sampling the profile contour from a thresholded image. For thresholding of the input profile image, the "Value" of the HSV model is exploited. No facial hair or glasses are allowed.

Kobayashi and Hara [42] apply an analytic approach to face detection. They are using a CCD camera in monochrome mode to obtain brightness distribution data of the human face. First, "base" brightness distribution was calculated as an average of brightness distribution data obtained from 10 subjects. Then, the system extracts the position of the irises by utilizing a crosscorrelation technique on the "base" data and the currently examined data. Once the irises are identified, the overall location of the face is determined by using relative locations of the facial features in the face. The observed subject should face the camera while siting at approximately 1m distance in front of it.

Yoneyama et al. [104] use an analytic approach to face detection too. The outer corners of the eyes, the height of the eyes, and the height of the mouth are extracted in an automatic way. Once these features are identified, the size of the examined facial area is normalized and an 8×10 rectangular grid is placed over the image. It is not stated which method has been applied and no limitation of the used method has been reported by Yoneyama et al.

Kimura and Yachida [41] utilize a Potential Net for face representation. An input image is normalized first by using the centers of the eyes and the center of the mouth tracked by the method proposed by Wu et al. [99]. This algorithm applies an integral projection method, which synthesizes the color and the edge information. Then, the Potential Net is fitted to the normalized image to model the face and its movement. The face should be without facial hair and glasses and in a direct face-to-face position with the camera [99].

4.1.2 Face Detection in Arbitrary Images

Two of the works surveyed in this paper perform automatic face detection in an arbitrary scene. Hong et al. [30] utilize the PersonSpotter system [81] in order to perform a real-time tracking of the head. The box bounding the head is used then as the image to which an initial labeled graph is fitted. The head-tracking module of the PersonSpotter system applies first a spatio-temporal filtering of the input image sequence. Then, a stereo algorithm determines the stereo disparities of the pixels that have been changed due to the movement. By inspecting the local maximums of the disparity histogram, image regions confined to a certain disparity interval are selected. The skin color detector and the convex region detector are applied to those regions. The bounding boxes confining the clusters of the outputs of both detectors are likely to correspond to heads and hands. For the case that the person is not moving any longer, the head-tracking module memorizes the last position of the person. The estimation of the current position and velocity of the head is achieved by using a linear predictive filter. Steffens et al. [81] reported that their system performs well in the presence of background motion, but fails in the case of covered or too much rotated faces.

Essa and Pentland [24] use the eigenspace method of Pentland et al. [65] to locate faces in an arbitrary scene. The method employs eigenfaces approximated using Principal Component Analysis (PCA) on a sample of 128 facial images. The eigenfaces define the subspace of sample images, i.e., so-called "face space" [91]. To detect the presence of a face in a single image, the distance of the observed image from the face space is calculated using the projection coefficients and the signal energy. To detect the presence of faces in an image sequence, a spatio-temporal filtering is performed, the filtered image is thresholded in order to analyze "motion blobs," and each motion blob that can represent a human head is then evaluated as a single image. The method of Pentland et al. [65] is real-time and has been successfully tested on a database of 7,562 images of some 3,000 people of both sexes, ranged in age and ethnicity, having varying head positions, headwear, and facial hair.

4.2 Facial Expression Data Extraction

After the presence of a face is detected in the observed scene, the next step is to extract the information about the shown facial expression. Both the applied face representation and the kind of input images affect the choice of the approach to facial expression data extraction.

In general, three types of face representation are mainly used in facial expression analysis: holistic (e.g., isodensity maps [38]), analytic (e.g., deformable templates [105]), and hybrid (e.g., analytic-to-holistic approach [45]). The face representations used by the surveyed systems are listed in Table 5. Depending on the face model a template-based or a feature-based method is applied for facial expression data extraction. Template-based methods fit a holistic face model to the input image or track it in the input image sequence. Feature-based methods localize the features of an analytic face model in the input image or track them in the input sequence. The methods utilized by the surveyed systems are listed in Table 6.

TABLE 5
The Utilized Face Models

Reference	Model
Holistic approach	
Edwards [18]	AAM
Hong [30]	Labeled graph
Huang [32]	PDM
Padgett [61]	Random block eigenvectors
Black [2]	Optical flow (in facial regions)
Otsuka [60]	Optical flow (in facial regions)
Analytic approach	
Hara [42]	FCPs model and 13 vertical lines
Pantic [62]	Dual-view point-based model
Zhao [107]	Frontal-view point-based model
Cohn [10]	Optical flow (facial points)
Hybrid approach	
Lyons [51]	Fiducial grid & Gabor wavelets
Yoneyama [104]	8x10 quadratic grid
Zhang [106]	Fiducial points & Gabor wavelets
Essa [24]	Optical flow (whole face)
Kimura [41]	Potential Net
Wang [95]	Labeled graph

4.2.1 Facial Data Extraction from Static Images: Template-Based Methods

As shown in Table 3, several surveyed systems can be classified as methods for facial expression analysis from static images. A first category of these utilizes a holistic or a hybrid approach to face representation (Table 5) and applies a template-based method for facial expression information extraction from an input image.

Edwards et al. [18] utilize a holistic face representation, which they refer to as the Active Appearance Model (AAM). To build their model they used facial images that were manually labeled with 122 points localized around the facial features. To generate a statistical model of shape variation, Edwards et al. aligned all training images into a common coordinate frame and applied PCA to get a mean shape. To build a statistical model of gray-level appearance, they warped each training image, by using a triangulation algorithm, so that its control points match the mean shape. By applying PCA to the gray-level information extracted from the warped images, they obtained a mean normalized gray-level vector. By applying PCA once more, an 80D vector of appearance parameters controlling both, the shape and the gray-levels of the model, has been obtained. To fit the AAM to an input image, Edwards et al. apply an AAM search algorithm which implies two stages [12]. In the training stage, for each of 88 training images labeled with 122 landmark points, known model displacements are applied and the corresponding difference vector is recorded. After the training data are generated, a multivariate multiple regression analysis is applied to model the relationship between the model displacement and the

TABLE 6
The Methods for Automatic Facial Expression Data Extraction

Reference	Method	Comment
Analysis from static facial images		
Template-based methods		
Edwards [18]	A multivariate multiple regression for modeling the relationship between the AAM displacement and the image difference and in the recognition phase to match the AAM to the input image.	Direct frontal view Faces without facial hair, glasses Hand labeling of the images
Hong [30]	Fitting a labeled graph (Fig. 2) to an input facial image by utilizing the method of elastic graph matching (Wiskott [98]).	Faces without facial hair, glasses Slightly rotated faces allowed Real-time process
Huang [32]	Fitting the PDM (Fig. 3-4) by applying a gradient-descent-based shape parameters estimation; fitting 3 parabolas to the mouth by applying gradient-based edge detector.	Direct frontal view Faces without facial hair, glasses No variation of the background
Yoneyama [104]	Gradient-based optical flow algorithm [31] for estimating an averaged optical flow in 80 20×20 pixels regions of the grid placed over a normalized image.	Direct frontal view Faces without facial hair, glasses Averaging the flow (drawback) Horizontal movement isn't modeled
Feature-based methods		
Hara [42]	Extracting the brightness distribution data along the 13 vertical facial lines; CCD camera in monochrome mode used.	Direct frontal view Faces without facial hair, glasses Horizontal movement isn't modeled Real-time process
Pantic [62]	Multiple feature detectors are applied per facial feature. From the localized contours of the prominent facial features the model features (Fig. 7) are extracted.	Dual view images Faces without facial hair, glasses 2 cameras mounted on user's head
Analysis from facial image sequences		
Template-based methods		
Black [2]	Robust regression scheme based on a brightness constancy assumption for image motion local models' parameter recovering. Coarse-to-fine gradient-based optical flow algorithm for estimating large motions.	Rigid head motions allowed Variations in lightning allowed The initial regions of the head and features are selected by hand
Otsuka [60]	Adapted gradient-based optical flow algorithm [31], [3], for estimating the image motion in the local areas of the right eye and mouth (Fig. 9).	Faces without facial hair, glasses Camera mounted on user's head Left eye motion isn't tracked
Essa [24]	Optical flow method of Simoncelli [80]: multi-scale coarse-to-fine Kalman filter for obtaining "noise-free" 2D motion field for a normalized facial image (Fig. 13).	Direct frontal view Faces with facial hair, glasses Variations in lightning allowed
Kimura [41]	Fitting a Potential Net (Fig. 10) to a normalized facial image by applying a differential and a Gaussian filter.	Direct frontal view Faces without facial hair, glasses Variations in lightning allowed 1st frame - expressionless face
Wang [95]	Fitting a labeled graph of 19 facial points (Fig. 11) by applying the method [7]: the minimization of the cost function is based on a simulated annealing procedure.	Direct frontal view Faces without facial hair, glasses Hand labeling of the 1st frame
Feature-based methods		
Cohn [10]	Hierarchical optical flow algorithm of Lucas and Kanade [49] for estimating the optical flow in 13×13 pixels regions surrounding the facial landmarks.	Direct frontal view Faces without facial hair, glasses 1st frame - expressionless face Manual normalization of frames Hand labeling of the 1st frame

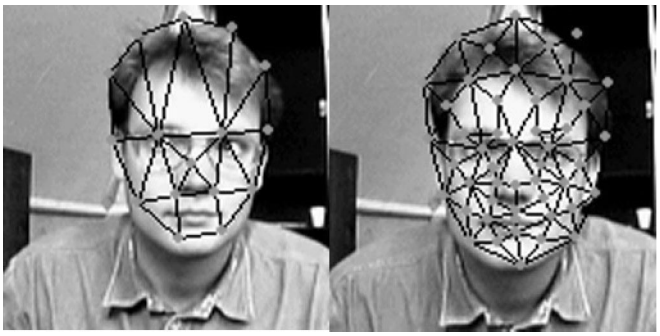


Fig. 2. A small model-graph (small GFK) and a dense model-graph (big GFK) [81].

image difference. In the recognition stage, the learned regression model is used to determine the movement of the face model. The AAM search algorithm has been tested on 100 hand labeled face images. Of these, 19.2 percent failed to converge to a satisfactory result [12]. The method works with images of faces without facial hair and glasses, which are hand-labeled with the landmark points beforehand approximated with the proposed AAM.

Hong et al. [30] utilize a labeled graph to represent the face. Each node of the graph consists of an array, which is called *jet*. Each component of a jet is the filter response of a certain *Gabor wavelet* [50] extracted at a point of the input image. Hong et al. use wavelets of five different frequencies and eight different orientations. They defined two different labeled graphs, called General Face Knowledge (GFK). A "big" GFK is a labeled graph with 50 nodes (Fig. 2), where to each node a 40-component jet of the corresponding landmark extracted from 25 individual faces has been assigned. A "small" GFK is a labeled graph with 16 nodes (Fig. 2). Each node contains a 12-component jet (four wave field orientations and three frequencies) that has been extracted from a set of eight faces. The small GFK is used to find the exact face location in an input facial image and the big GFK is used to localize the facial features. Hong et al. utilize the PersonSpotter system [81] and the method of elastic graph matching proposed by Wiskott [98] to fit the model-graph to a surface image. First, the small GFK is moved and scaled over the input image until a place of the best match is found. After the matching is performed, the exact face position is derived from the canonical graph size value (the mean Euclidean distance of all nodes from the center of gravity). Then the big GFK is fitted to the cropped face region and a node-weighting method is applied. A low weight is assigned to the nodes on the face and hair boundary and a high weight is assigned to the nodes on the facial features. The big GFK with weighted nodes is used further to emotionally classify the shown facial expression. Although Hong et al. utilize the PersonSpotter system [81], which deals with real-time processing of video sequences, they perform facial expression analysis from static images. The dense model-graph seems very suitable for facial action coding based on the extracted deformations of the graph. However, this issue has not been discussed by Hong et al. [30].

To represent the face, Huang and Huang [32] utilize a point distribution model (PDM), [11]. The used PDM has been generated from 90 facial feature points that have been manually localized in 90 images of 15 Chinese subjects showing six basic emotions (Figs. 3 and 4). The mouth is included in the model by approximating the contour of the mouth with three parabolic curves. Since the proposed model is a combination of the PDM and a mouth template, it is arguably as close to a feature-based model as to a template-based model. We classified it as a holistic face model since the PDM models the face as a whole and interacts with the estimated face region of an input image as entire. After an initial placement of the PDM in the input image (Section 4.1.), the method of Huang and Huang moves and deforms the entire PDM simultaneously. Here, a gradient-based shape parameters estimation, which minimizes the overall gray-level model fitness measure, is applied. The search for the mouth starts by defining an appropriate search region on basis of the fitted PDM. Then the darkest point in each vertical strip of the search region is found. A gray-level thresholding is applied to eliminate the misleading points and a parabolic curve is used to approximate the mouth-through line. The edges with the strongest gradient, located above this line, are approximated with another parabolic curve to represent the upper lip. The same method is applied to find the lower lip. The method will not work out if the teeth are visible, i.e., if there is no dark region between the lips. Also, successfulness of the method is strongly constrained (Section 4.1.1).

Fig. 3. Aligned training set for generation of PDM model (reprinted from [32] with permission from Academic Press) © 1997 Academic Press.

Fig. 4. Fitted PDM model (reprinted from [32] with permission from Academic Press) © 1997 Academic Press.

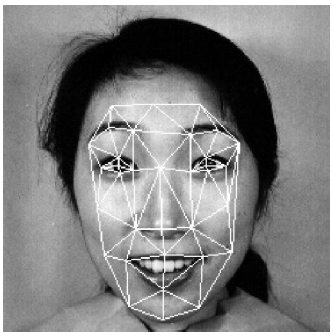


Fig. 5. Fiducial grid of facial points ([51]).

Padgett and Cottrell [61] also use a holistic face representation, but they do not deal with facial expression information extraction in an automatic way. They made use of the facial emotion database assembled by Ekman and Friesen [20], [21], digitized 97 images of six basic emotional facial expressions, and scaled them so that the prominent facial features were located in the same image region. Then, in each image, the area around each eye was divided into two vertically overlapping 32×32 pixel blocks and the area around the mouth was divided into three horizontally overlapping 32×32 pixel blocks. PCA of 32×32 pixel blocks randomly taken over the entire image was applied in order to generate the eigenvectors. The input to a NN used for emotional classification of an expression was the normalized projection of the seven extracted blocks on the top 15 principal components.

Yoneyama et al. [104] use a hybrid approach to face representation. They fit an 8×10 quadratic grid to a normalized facial image (see Section 4.1.1). Then, an averaged optical flow is calculated in each of the 8×10 regions. To calculate the optical flow between a neutral and an examined facial expression image, they use the optical flow algorithm proposed by Horn and Schunck [31]. The magnitude and the direction of the calculated optical flows are simplified to a ternary value magnitude in only the vertical direction. The information about a horizontal movement is excluded. Hence, the method will fail to recognize any facial appearance change that involves a horizontal movement of the facial features. The face should be without facial hair and glasses and no rigid head motion may be encountered for the method to work correctly.

Zhang et al. [106] use a hybrid approach to face representation, but do not deal with facial expression information extraction in an automatic way. They use 34 facial points (Fig. 5) for which a set of Gabor wavelet coefficients is extracted. Wavelets of three spatial frequencies and six orientations have been utilized. Zhang et al. deal only with 256×256 pixels frontal view images of nine female Japanese subjects, manually normalized so that the distance between the eyes is 60 pixels. A similar face representation was recently used by Lyons et al. [51] for expression classification into the six basic plus "neutral" emotion categories. They used a fiducial grid of manually positioned 34 nodes on 256×256 pixels images used in [106], but apply wavelets of five spatial frequencies and six angular orientations.

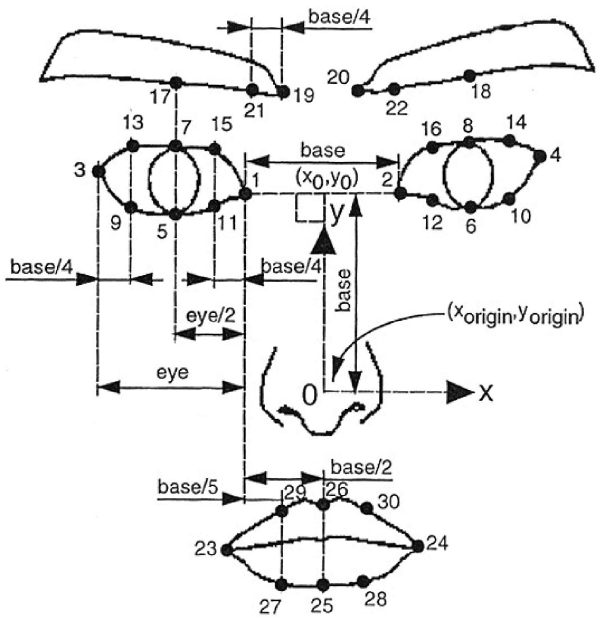


Fig. 6. Facial Characteristic Points ([42]).

4.2.2 Facial Data Extraction from Static Images: Feature-Based Methods

The second category of the surveyed methods for automatic facial expression analysis from static images uses an analytic approach to face representation (Table 3, Table 5) and applies a feature-based method for expression information extraction from an input image.

In their earlier work [43], [44], Kobayashi and Hara proposed a geometric face model of 30 FCPs (Fig. 6). In their later work [42], they utilize a CCD camera in monochrome mode to obtain a set of brightness distributions of 13 vertical lines crossing the FCPs. First, they normalize an input image by using an affine transformation so that the distance between irises becomes 20 pixels. From the distance between the irises, the length of the vertical lines is empirically determined. The range of the acquired brightness distributions is normalized to [0,1] and these data are given further to a trained NN for expression emotional classification. A shortcoming of the proposed face representation is that the facial appearance changes encountered in a horizontal direction cannot be modeled. The real-time system developed by Kobayashi and Hara works with online taken images of subjects with no facial hair or glasses facing the camera while siting at approximately 1m distance from it.

Pantic and Rothkrantz [62] are utilizing a point-based model composed of two 2D facial views, the frontal and the side view. The frontal-view face model is composed of 30 features. From these, 25 features are defined in correspondence with a set of 19 facial points (Fig. 7) and the rest are some specific shapes of the mouth and chin. The utilized side-view face model consists of 10 profile points, which correspond with the peaks and valleys of the curvature of the profile contour function (Fig. 7). To localize the contours of the prominent facial features and then extract the model features in an input dual-view, Pantic and Rothkrantz apply multiple feature detectors for each

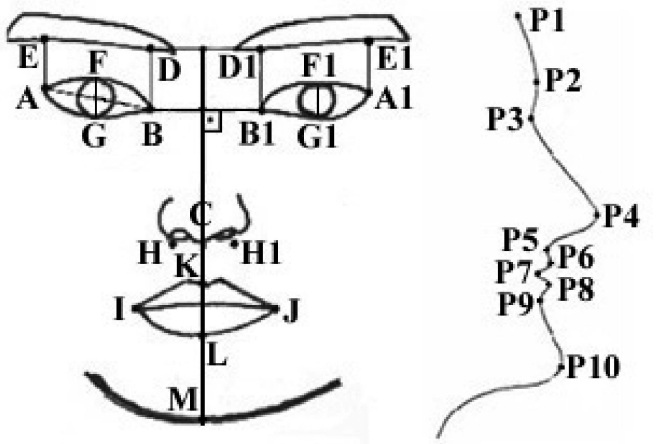


Fig. 7. Facial points of the frontal-view face model and the side-view face model ([63]).

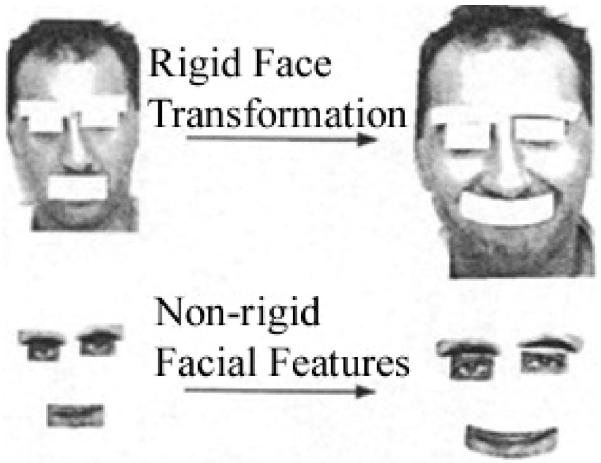


Fig. 8. Planar model for representing rigid face motions and affine-plus-curvature model for representing nonrigid facial motions ([3]).

prominent facial feature (eyebrows, eyes, nose, mouth, and profile). For example, to localize the eyes they use methods proposed in [94], and [37] with [96]. Then, the best of the acquired (redundant) results is chosen. This is done based on both, the knowledge about the facial anatomy (used to check the correctness of the result of a certain detector) and the confidence in the performance of a specific detector (assigned to it based on its testing results). The performance of the detection scheme was tested on 496 dual views. Human observers in 89 percent approved when visually inspected the achieved localization of the facial features. The system cannot deal with minor inaccuracies of the extracted facial data and it deals merely with images of faces without facial hair or glasses.

Zhao et al. [107] also utilize a point-based frontal-view face model but do not deal with automatic facial expression data extraction. They utilize 10 facial distances, to manually measure 94 images selected from the facial emotion database assembled by Ekman and Friesen [20], [21]. These data are used further for expression emotional classification.

4.2.3 Facial Data Extraction from Image Sequences: Template-Based Methods

A first category of the surveyed approaches to automatic facial expression analysis from image sequences uses a holistic or a hybrid approach to face representation (Table 3, Table 5) and applies a template-based method for facial expression information extraction from an input image sequence.

Black and Yacoob [2] are using local parameterized models of image motion for facial expression analysis. They utilize an affine-, a planar-, and an affine-plus-curvature flow model. The planar model is used to represent rigid facial motions. The motion of the plane is used to stabilize two frames of the examined image sequence and the motions of the facial features are then estimated relatively to the stabilized face. Nonrigid motions of facial features within the local facial areas of the eyebrows, eyes, and mouth (Fig. 8) are represented by affine-plus-curvature model. To recover the parameters of the flow models, a robust regression scheme based on the brightness constancy assumption is employed. To cope with large motions, a coarse-to-fine gradient-descent strategy is used. In the approach proposed by Black and Yacoob, the initial regions for the head and the facial features were selected by hand and thereafter automatically tracked.

By applying an adapted gradient-based optical flow algorithm [3], Otsuka and Ohya [59] are estimating the motion in the local facial areas of the right eye and the mouth (Fig. 9). The input facial images are taken by a camera mounted on a helmet worn by the subject and subsampled by eight in both directions [59]. After the optical-flow algorithm is applied, a 2D Fourier transform is utilized to the horizontal and the vertical velocity field and the lower-frequency coefficients are extracted as a 15D feature vector, which is used further for facial expression emotional classification. Otsuka and Ohya are taking advantage of the face symmetry when estimating the motion just in the local facial areas of the right eye and the mouth. As a consequence, their method is not sensitive to unilateral appearance changes of the left eye.

Essa and Pentland [24] are utilizing a hybrid approach to face representation. First, they applied the eigenspace method [65] to automatically track the face in the scene (Section 4.1.2) and extract the positions of the eyes, nose, and mouth. The method for extracting the prominent facial features employs eigenfeatures approximated using PCA on a sample of 128 images. The eigenfeatures define a so-called "feature space." To detect the location of the prominent facial features in a given image, the distance of each feature-image from the relevant feature space is computed using a FFT and a local energy computation. The extracted position of the prominent facial features is further used to normalize the input image. A 2D spatio-temporal motion energy representation of facial motion estimated from two consecutive normalized frames is used as a dynamic face model. Essa and Pentland use the optical flow computation method proposed by Simoncelli [80]. This approach uses a multiscale coarse-to-fine Kalman filter to obtain motion estimates and error-covariance information. The method computes first a mean velocity vector, which represents the estimated flow from consecutive normalized facial images of a video sequence. The flow covariances between different frames are stored and used together with

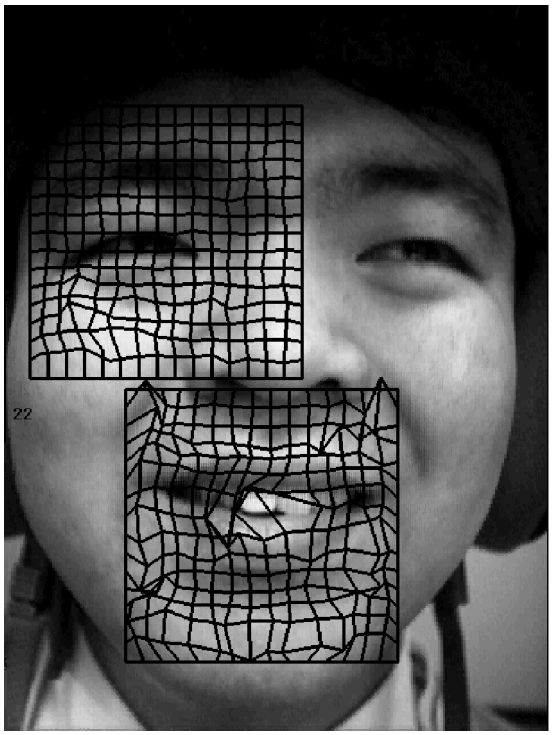


Fig. 9. Motion vector field represented in the deformation of two grids ([60]).

the recursive continuous time Kalman filter to calculate the error predictions, based on previous data, and to obtain a "corrected," "noise-free" 2D motion field. The method has been applied to frontal-view facial image sequences.

Another system, where a hybrid approach to face representation is utilized, is proposed by Kimura and Yachida [41]. They are utilizing a Potential Net. To fit the Potential Net to a normalized facial image (see Section 4.1.1), they compute first the edge image by applying a differential filter. Then, in order to extract the external force, which is a smooth gradient of the edge image, they are applying a Gaussian filter. The filtered image is referred to as a "potential field" to which the elastic net model (Fig. 10) is placed. The net deforms further governed by the elastic force of the potential field. The method seems suitable for facial action encoding based on the extracted deformations of the net. However, Kimura and Yachida have not discussed this issue.

Wang et al. [95] use a hybrid approach to face representation too. They utilize 19 facial feature points (FFPs)—seven FFPs to preserve the local topology and 12 FEFPs (depicted as • in Fig. 11) for facial expression recognition. The FFPs are treated as nodes of a labeled graph that are interconnected with links representing the Euclidean distance between the nodes. The links are weighted with empirically set parameters denoting some properties of the facial features to which the FFPs belong. For example, the links between the mouth nodes are weighted with a smaller weight since the mouth can deform violently. The initial location of the FFPs in the first frame of an input image sequence is assumed to be known. To track the FFPs in the rest of the frames, Wang et al. use a system that consists of two layers, a memory layer

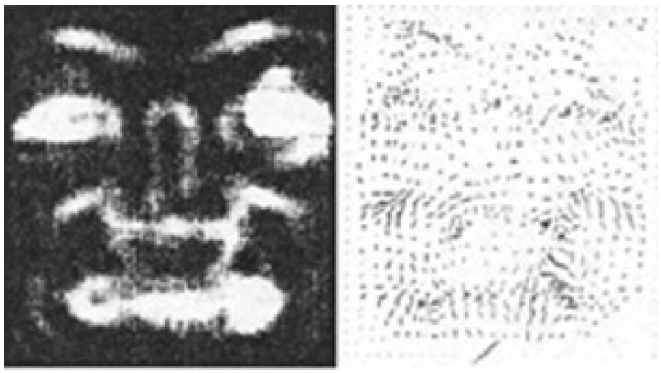


Fig. 10. Potential Field and corresponding Potential Net ([41]).

and an input layer. The correspondence between the FFPs tracked in two consecutive frames is treated as a labeled graph matching problem as proposed by Buhmann et al. [7], where the antecedent frame is treated as the memory layer and the current frame as the input layer. The graph matching is realized as a dynamic process of node diffusion, which minimizes a cost function based on a simulated annealing procedure. The observed faces should be without facial hair or glasses, no rigid head motion may be encountered, the first frame of the examined image sequence should represent an expressionless face and the FFPs should be marked in the first frame for the method to work correctly. The face model used by Wang et al. represents a way of improving the labeled-graph-based models (e.g., [30]) to include intensity measurement of the encountered facial expressions based on the information stored in the links between the nodes.

4.2.4 *Facial Data Extraction from Image Sequences: Feature-Based Methods*

Only one of the surveyed methods for automatic facial expression analysis from image sequences utilizes an analytic face representation (Table 3, Table 5) and applies a feature-based method for facial expression information extraction. Cohn et al. [10] use a model of facial landmark points localized around the facial features, hand-marked with a mouse device in the first frame of an examined image sequence. In the rest of the frames, a hierarchical optical flow method [49] is used to track the optical flows of 13×13 windows surrounding the landmark points. The displacement of each landmark point is calculated by subtracting its normalized position in the first frame from its current normalized position (all frames of an input

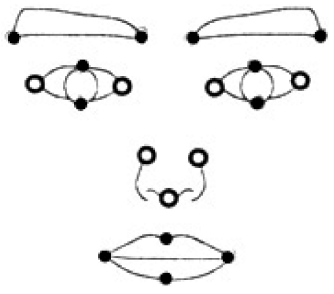


Fig. 11. The FFPs ([95]).

TABLE 7
Facial Expression Classification in Terms of Facial Actions

Reference	Method	# Facial actions	Test cases	Accuracy
Analysis from static facial images				
Rule-based methods				
Pantic [62]	Expert System rules	31 facial actions	496 dual views 8 subjects	89%
Analysis from facial image sequences				
Template-based methods				
Cohn [10]	Discriminant functions	8 AUs + 7 AUs combinations	504 sequences 100 subjects	88%
Essa [24]	Spatio-temporal motion-energy templates	2 facial actions	22 sequences 8 subjects	100%
Rule-based methods				
Black [2]	Thresholded motion parameters in a rule-form	-	70 sequences 40 subjects	88%

sequence are manually normalized). The displacement vectors, calculated between the initial and the peak frame, represent the facial information used for recognition of the displayed facial actions. The face should be without facial hair/ glasses, no rigid head motion may be encountered, the first frame should be an expressionless face, and the facial landmark points should be marked in the first frame for the method to work correctly.

4.3 Facial Expression Classification

The last step of facial expression analysis is to classify (identify, interpret) the facial display conveyed by the face. The surveyed facial expression analyzers classify the encountered expression (i.e., the extracted facial expression information) either as a particular facial action or a particular basic emotion. Some of the systems perform both. Independent of the used classification categories, the mechanism of classification applied by a particular surveyed expression analyzer is either a template-based- or a neural-network-based- or a rule-based- classification method. The applied methods for expression classification in terms of facial actions are summarized in Table 7. Table 8 summarizes the utilized methods for facial expression emotional classification.

If a template-based classification method is applied, the encountered facial expression is compared to the templates defined for each expression category. The best match decides the category of the shown expression. In general, it is difficult to achieve a template-based quantified recognition of a nonprototypic facial expression. There are infinitely a lot of combinations of different facial actions and their intensities that should be modeled with a finite set of templates. The problem becomes even more difficult due to the fact that everybody has his/her own maximal intensity of displaying a certain facial action.

Although the neural networks represent a " black box" approach and arguably could be, classified as template-based methods, we are classifying the neural-network-based methods separately. We are doing so because a typical neural network can perform a quantified facial expression categorization into multiple classes while, in general, the template-based methods cannot achieve such a performance. In a neural-network-based classification approach, a facial expression is classified according to the categorization process that the network "learned" during a training phase. Most of the neural-network-based classification methods utilized by the surveyed systems perform facial expression classification into a single category. Recognition of nonprototypic facial expressions is feasible, however, if each neural network output unit is associated with a weight from the interval [0,1], as proposed in [106], [44], [71], [56], instead of being associated with either 0 or 1 (e.g., [42], [107]). As it can be seen from Table 8, we classified some of the expression classifiers as template-based methods although they utilize a neural network (i.e., Yonoyama et al. [104]). We are doing so because the overall characteristics of these methods fit better the overall properties of the template-based expression classification approaches.

The rule-based classification methods, utilized by the surveyed systems, classify the examined facial expression into the basic emotion categories based on the previously encoded facial actions (Table 7, Table 8). The prototypic expressions, which characterize the emotion categories, are first described in terms of facial actions. Then, the shown expression, described in terms of facial actions, is compared to the prototypic expressions defined for each of the emotion categories and classified in the optimal fitting category.

4.3.1 Expression Classification from Static Images: Template-Based Methods

A first category of the surveyed methods for automatic expression analysis from static images applies a template-based method for expression classification. The methods in this category perform expression classification into a single basic emotion category.

Given a new example of a face and the extracted parameters of AAM (Section 4.2.1), the main aim of Edwards et al. [18] is to identify the observed individual in a way which is invariant to confounding factors such as pose and facial expression. To achieve this goal, they

TABLE 8
The Methods for Facial Expression Emotional Classification

Reference	Method	#	Test cases	Accuracy
Analysis from static facial images				
Template-based methods				
Edwards [18]	PCA based on Mahalonobis distance [27] and LDA	7	200 images 25 subjects	74%
Hong [30]	Personalised galleries and Elastic graph matching [98]	7	>175 images 25 subjects	81%
Huang [32]	2D emotion space (PCA) & minimum distance classifier	6	90 images 15 subjects	84.5%
Lyons [51]	PCA and LDA of the labelled-graph vectors	7	193 images 9 Japanese females	75 - 92%
Yoneyama [104]	Two 14x14 Hopfield NNs with learning [36]	4	-	-
Neural-network-based methods				
Hara [42]	234x50x6 NN with backpropagation learning	6	90 images 15 subjects	85%
Padgett [61]	15x10x7 NN with backpropagation learning	7	84 Ekman's photos	86%
Zhang [106]	646x7x7 NN with RPROP propagation [72]	7	213 images 9 Japanese females	90%
Zhao [107]	10x10x3 NN with backpropagation learning	6	94 Ekman's photos	100%
Rule-based methods				
Pantic [62]	Expert System rules	6	265 dual views 8 subjects	91%
Analysis from facial image sequences				
Template-based methods				
Essa [24]	Spatio-temporal motion-energy templates (Fig. 13)	4	30 sequences 8 subjects	98%
Kimura [41]	3D emotion space (PCA)	3	-	-
Otsuka [60]	HMM & Baum-Welch training method	6	-	-
Wang [95]	Averaged B-splines of feature trajectories & method [69] for distance minimization	3	29 sequences 8 subjects	95%
Rule-based methods				
Black [2]	Temporal consistency of the mid-level predicates which describe the motion of the facial features	6	70 sequences 40 subjects	88%

utilized the Mahalonobis distance measure [27] on a representative set of training facial images. This classifier assumes that the intraclass variation (pose and expression) is very similar for each individual. Edwards et al. used Linear Discriminant Analysis (LDA) to separate linearly the interclass variability (identity) from the intraclass variability. They showed that a subspace could be constructed that is orthogonal to matrix D (a matrix of orthogonal vectors describing the principal types of interclass variation), which models only the intraclass variation due to change in pose, expression, and lightning. The expression recognition performance of the AAM has been trained and tested on 2×200 images of six basic emotional expressions shown by 25 subjects. The images were chosen for limited pose and lightning variation. The achieved recognition rate for the six basic and "neutral" emotion categories was 74 percent. Edwards et al. explain the low recognition rate by the limitations and unsuitability of the utilized linear classifier [18]. It is not known how the method will behave in the case of an unknown subject.

To achieve expression classification into one of the six basic plus "neutral" emotion categories, Hong et al. [30] made the assumption that two persons who look alike have a similar way for showing the same expression. First, they fit a labeled graph (Fig. 2) to an input facial image (Section 4.2.1). Then, the best matching person, whose personalized gallery is available, is found by applying the method of elastic graph matching proposed by Wiskot [98]. The personalized galleries of nine people have been utilized, where each gallery contained 28 images (four images per expression). The personalized gallery of the best matching person is used to make the judgement on the category of the observed expression. The method has been tested on images of 25 subjects. The achieved recognition rate was 89 percent in the case of the familiar subjects and 73 percent in the case of unknown persons. As indicated by Hong et al., the availability of the personalized galleries of more individuals would probably increase the system's performance. The time necessary for performing a full analysis of an incoming facial image is about eight seconds.

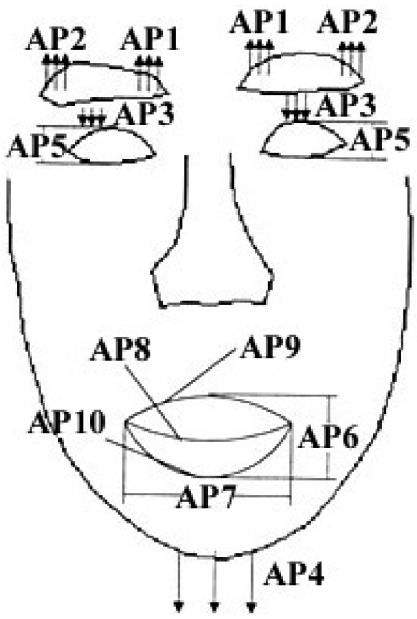


Fig. 12. APs (reprinted from [32] with permission from Academic Press) © 1997 Academic Press.

In order to perform emotional classification of the observed facial expression, Huang and Huang [32] perform an intermediate step by calculating 10 Action Parameters (APs, Fig. 12). The difference between the model feature parameters (Figs. 3 and 4) found in an expressionless face and those found in the examined facial expression of the same person generates the APs. Experimentally, they found out that the first two terms of eigenvalues could represent more than 90 percent of APs variations. They used a minimum distance classifier to cluster the two principal action parameters of 90 training image samples into six clusters representing six basic emotional expressions. Since the principal component distribution of each expression is overlapped with the distribution of at least two other expressions, three best matches are selected. The highest score of the three correlation determines the final classification of the examined expression. The proposed method has been tested on another 90 images shown by the same subjects. The achieved correct recognition ratio was 84.5 percent. It is not known how the method will behave in the case of unknown subjects. Also, the descriptions of the emotional expressions, given in terms of facial actions, are incomplete. For example, an expression with lowered mouth corners and raised eyebrows will be classified as sadness.

On a fiducial grid of manually positioned 34 nodes (Fig. 5), Lyons et al. [51] sample the amplitude of the complex valued Gabor transform coefficients and combine these data into a single vector which they call labeled-graph vector (LG vector). The ensemble of LG vectors from a training set of images is further subjected to PCA. The ensemble of LG-PCA vectors from the training set is then analyzed using LDA in order to separate vectors into clusters having different facial attributes. Lyons et al. experimented with the binary classifiers for the presence or absence of a particular facial expression. They built six binary classifiers, one for each basic emotion category, and combined them into a single facial expression classifier. An input LG vector is classified by being projected along the discriminant vectors calculated for each independently trained binary classifier. For an input image that is positively classified for two or more emotion categories, the normalized distances to the cluster centers are used as a deciding factor. The input sample is classified as a member of the nearest cluster. An input image that is not positively classified for any category is classified as "neutral." To test their method, Lyons et al. used a set of 193 images of different facial expressions displayed by nine Japanese females, which has been used by Zhang et al. [106] (see Section 4.3.2). The entire set of images was divided into 10 segments; the discriminant vectors were calculated using nine of these segments and the generalization performance was tested on the remaining segment; the results were averaged over all 10 distinct partitions. The generalization rate was 92 percent. The method was also tested on this image set only partitioned into nine segments, each corresponding to one expresser. The generalization rate was 75 percent for recognition of expression of a novel subject.

Yoneyama et al. [104] extract 80 facial movement parameters, which describe the change between an expressionless face and the currently examined facial expression of the same subject (Section 4.2.1). To recognize four types of expressions (sadness, surprise, anger, and happiness), they use 2 bits to represent the values of 80 parameters ("1 -1" for up, "-1 1" for down and "-1 -1" for no movement) and two identical discrete Hopfield networks. Each network consists of 14×14 neurons, where 36 neurons are added to form a neurons square and have -1 as an initial value. The first net (NN1) is trained on 40 data representing four expressions shown by 10 subjects. NN2 is trained just on four data representing "the most clearly shown" four expressions. The NNs were trained using the Personnaz learning rule [36]. For each examined image, the output of the NN1 is matched with all of the examples used for training of the NN1 and the Euclidean distances are calculated. The distances are averaged per expression. If the difference between the minimal average and the second minimal average is greater than 1, the category of the examined expression is decided. Otherwise, the output of the NN2 is matched to the examples used for training of the NN2 in order to decide a final category of the shown expression. The average recognition rate was 92 percent. The images used for training of the networks were also used for their testing.

4.3.2 Expression Classification from Static Images: Neural-Network-Based Methods

A second category of the surveyed methods for automatic facial expression analysis from static images applies a neural network for facial expression classification. Except the method proposed by Zhang et al. [106], the methods belonging to this category perform facial expression classification into a single basic emotion category.

For classification of expression into one of six basic emotion categories, Hara and Kobayashi [42] apply a $234 \times 50 \times 6$ back-propagation neural network. The units of the input layer correspond to the number of the brightness distribution data extracted from an input facial image (Section 4.2.2) while each unit of the output layer

corresponds to one emotion category. The neural network has been trained on 90 images of six basic facial expressions shown by 15 subjects and it has been tested on a set of 90 facial expressions images shown by another 15 subjects. The average recognition rate was 85 percent. The process takes 66.7ms.

For emotional classification of an input facial image into one of 6 basic plus "neutral" emotion categories, Padgett and Cottrell [61] utilize a back-propagation neural network. The input to the network consists of the normalized projection of seven 32×32 pixel blocks on the first 15 principal components of previously generated random-blocks-eigenspace (Section 4.2.1). The hidden layer of the NN contains 10 nodes and employs a nonlinear Sigmoid activation function. The output layer of the NN contains seven units, each of which corresponds to one emotion category. Padgett and Cottrell used the images of six basic plus "neutral" expressions shown by 12 subjects. They trained the network on the images of 11 subjects and tested it on the images of the 12th subject. By changing the training and the testing set, they trained 12 networks. The average correct recognition rate achieved was 86 percent.

Zhang et al. [106] employ a $680 \times 7 \times 7$ neural network for facial expression classification into six basic plus "neutral" emotion categories. The input to the network consists of the geometric position of the 34 facial points (Fig. 11) and 18 Gabor wavelet coefficients sampled at each point. The neural network performs a nonlinear reduction of the input dimensionality and makes a statistical decision about the category of the observed expression. Each output unit gives an estimation of the probability of the examined expression belonging to the associated category. The network has been trained using a resilient propagation [72]. A set of 213 images of different expressions displayed by nine Japanese females has been used to train and test the used network. The database has been partitioned into ten segments. Nine segments have been used to train the network while the remaining segment has been used to test its recognition performance. This process has been repeated for each of the 10 segments and the results of all 10 trained networks have been averaged. The achieved recognition rate was 90.1 percent. The performance of the network is not tested for recognition of expression of a novel subject.

Zhao and Kearney [107] utilize a $10 \times 10 \times 3$ back-propagation neural network for facial expression classification into one of six basic emotion categories. They used 94 images of six basic facial expressions selected from the emotion database assembled by Ekman and Friesen [20], [21]. On each image, 10 distances were manually measured. The difference between a distance measured in an examined image and the same distance measured in an expressionless face of the same person was normalized. Then, each such measure was mapped into one of the eight signaled intervals of the appropriate standard deviation from the corresponding average. These intervals formed the input to the NN. The output of the NN represents the associated emotion (e.g., the string "001" is used to represent happiness). The NN was trained and tested on the whole set of data (94 images) with 100 percent recognition rate. It is not known how the method will behave in the case of an unknown subject.

4.3.3 Expression Classification from Static Images: Rule-Based Methods

Just one of the surveyed methods for automatic facial expression analysis from static images applies a rule-based approach to expression classification. The method proposed by Pantic and Rothkrantz [62] achieves automatic facial action coding from an input facial dual-view in few steps. First, a multidetector processing of the system performs automatic detection of the facial features in the examined facial image (Section 4.2.2). From the localized contours of the facial features, the model features (Fig. 7) are extracted. Then, the difference is calculated between the currently detected model features and the same features detected in an expressionless face of the same person. Based on the knowledge acquired from FACS [21], the production rules classify the calculated model deformation into the appropriate AUs-classes (total number of classes is 31). The performance of the system in automatic facial action coding from dual-view images has been tested on a set of 496 dual views (31 expressions of separate facial actions shown twice by eight human experts). The average recognition rate was 92 percent for the upper face AUs and 86 percent for the lower face AUs.

Classification of an input facial dual-view into multiple emotion categories is performed by comparing the AU-coded description of the shown facial expression to AU-coded descriptions of six basic emotional expressions, which have been acquired from the linguistic descriptions given by Ekman [22]. The classification into and, then, quantification of the resulting emotion labels is based on the assumption that each subexpression of a basic emotional expression has the same influence on scoring that emotion category. The overall performance of the system has been tested on a set of 265 dual facial views representing six basic and various blended emotional expressions shown by eight subjects. A correct recognition ratio of 91 percent has been reported. The dual-views used for testing of the system have been recorded under constant illumination and none of the subjects had a moustache, a beard, or wear glasses.

4.3.4 Expression Classification from Image Sequences: Template-Based Methods

The first category of the surveyed methods for automatic facial expression analysis from facial image sequences applies a template-based method for expression classification.

The facial action recognition method proposed by Cohn et al. [10] applies separate discriminant function analyzes within facial regions of the eyebrows, eyes, and mouth. Predictors were facial points displacements (Section 4.2.4) between the initial and peak frames in an input image sequence. Separate group variance-covariance matrices were used for classification. Image sequences (504) containing 872 facial actions displayed by 100 subjects have been used. The images have been recorded under constant illumination, using fixed light sources and none of the subjects wear glasses [48]. Data were randomly divided into training and test sets of image sequences. They used two discriminant functions for three facial actions of the eyebrow region, two discriminant functions for three facial actions of the eye region,

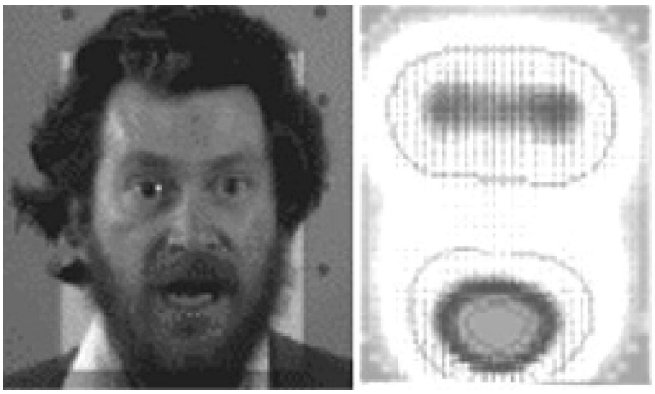


Fig. 13. The spatio-temporal template for surprise ([24]).

and five discriminant functions for nine facial actions of the nose and mouth region. The accuracy of the classification was 92 percent for the eyebrow region, 88 percent for the eye region and 83 percent for the nose/mouth region. The method proposed by Cohn et al. deals neither with image sequences containing several facial actions in a row, nor with inaccurate facial data, nor with facial action intensity (yet the concept of the method makes it possible).

Essa and Pentland [24] use a control-theoretic method to extract the spatio-temporal motion-energy representation of facial motion for an observed expression (Section 4.2.3). By learning "ideal" 2D motion views for each expression category, they generated the spatio-temporal templates (Fig. 13) for six different expressions—two facial actions (smile and raised eyebrows) and four emotional expressions (surprise, sadness, anger, and disgust). Each template has been delimited by averaging the patterns of motion generated by two subjects showing a certain expression. The Euclidean norm of the difference between the motion energy template and the observed image motion energy is used as a metric for measuring similarity/dissimilarity. When tested on 52 frontal-view image sequences of eight people showing six distinct expressions, a correct recognition rate of 98 percent has been achieved.

Kimura and Yachida [41] fit a Potential Net to each frame of the examined facial image sequence (Section 4.2.3, Fig. 10). The pattern of the deformed net is compared to the pattern extracted from an expressionless face (the first frame of a sequence) and the variation in the position of the net nodes is used for further processing. Kimura and Yachida built an emotion space by applying PCA on six image sequences of three expressions—anger, happiness, and surprise—shown by a single person gradually, from expressionless to a maximum. The eigenspace spanned by the first three principal components has been used as the emotion space, onto which an input image is projected for a quantified emotional classification. The proposed method has been unsuccessfully tested for image sequences of unknown subjects. A very small number of training examples (six sequences) and an insufficient diversity of the subjects (one person) have probably caused this.

Otsuka and Ohya [60] match the temporal sequence of the 15D feature vector (Section 4.2.3) to the models of the six basic facial expressions by using a left-to-right Hidden Markov Model. The used HMM consists of five states, namely, relaxed (S_1, S_5), contracted (S_2), apex (S_3), and relaxing (S_4). To facilitate recognition of a single image sequence, the transition from the final state to the initial state is added. To make recognition of multiple sequences of expression images feasible, the transition from a final state to the initial states of other categories is added. The transition probability and the output probability of each state are obtained from sample data by using the Baum-Welch algorithm. The initial probability is estimated by applying a k-means clustering algorithm on the sample data in which a squared sum of vector components is added as an extra component. The HMM was trained on 120 image sequences, shown by two male subjects. The method was tested on image sequences shown by the same subjects. Therefore, it is not known how the method will behave in the case of an unknown expresser. Although Otsuka and Ohya claim that the recognition performance was good, they do not define the extent of "good."

Wang et al. [95] utilize a 19-points labeled graph with weighted links to represent the face (Section 4.2.3, Fig. 11). Twelve of these points (FEFPs) are used for expression recognition. For each of three emotion categories—anger, happiness, and surprise—they use 12 average FEFP B-spline curves, one for each FEFP, to construct the expression model. Each curve describes the relationship between the expression change and the displacement of the corresponding FEFP. Each expression model has been defined from 10 image sequences displayed by five subjects. The category of an expression is decided by determining the minimal distance between the actual trajectory of FEFPs and the trajectories defined by the models. The distance functions are minimized using the method proposed by Brent [69]. The degree of expression change is determined based on the displacement of the FEFPs in the consecutive frames. The method has been tested on 29 image sequences of three emotional expressions shown by eight subjects (young and of Asian ethnicity). The images were acquired under constant illumination and none of the subjects had a moustache, a beard or wear glasses. The average recognition rate was 95 percent. The average process time was 2.5s/frame.

4.3.5 *Expression Classification from Image Sequences: Rule-Based Methods*

Just one of the surveyed methods for automatic facial expression analysis from image sequences applies a rule-based approach to expression classification. Black and Yacoob [2], [3], utilize local parameterized models of image motion to represent rigid head motions and nonrigid facial motions within the local facial areas (Fig. 8). The motion parameters (e.g., translation and divergence) are used to derive the midlevel predicates that describe the motion of the facial features. Each midlevel predicate is represented in a form of a rule, where the left part of the rule is a comparison of a motion parameter to a certain threshold and the right part of the rule is the derived predicate. The thresholds are dependent on the face size in the image and were set empirically from a few sequences. Black and Yacoob did not give a full list of the midlevel predicates and the number of different facial actions that the method can recognize is not known. In their method, the facial expression emotional classification considers the temporal consistency of the midlevel representation predicates. For

each of six basic emotional expressions, they developed a model represented by a set of rules for detecting the beginning and ending of the expression. The rules are applied to the predicates of the midlevel representation. The method has been tested on 70 image sequences containing 145 expressions shown by 40 subjects ranged in ethnicity and age. The expressions were displayed one at the time. The achieved recognition rate was 88 percent. "Lip biting" is sometimes mistakenly identified as a smile [2]. Also, the method does not deal with blends of emotional expressions. For example, a blend of the angry and scared expressions is recognized as disgust. The reason lies in the rules used for classification.

5 Discussion

We have explored and compared a number of different recently presented approaches to facial expression detection and classification in static images and image sequences. The investigation compared automatic expression information extraction using facial motion analysis [2], [10], [48], [24], [60], holistic spatial pattern analysis [18], [30], [32], [104], [41], [95], and analysis of facial features and their spatial arrangement [42], [62], [10]. This investigation also compared facial expression classification using holistic spatial analysis [18], [30], [32], [51], [104], [61], holistic spatio-temporal analysis [2], [24], [41], [60], [95], gray-level pattern analysis using local spatial filters [51], [106], and analytic (feature-based) spatial analysis [32], [42], [107], [62]. The number of the surveyed systems is rather large and the reader might be interested in the results of the performed comparison in terms of the best performances. Yet, we deliberately didn't make an attempt to label some of the surveyed systems as being better than some other systems presented in the literature. We believe that a well-defined and commonly used single database of testing images (image sequences) is the necessary prerequisite for "ranking" the performances of the proposed systems in an objective manner. Since such a single testing data set has not been established yet, we left the reader to decide the ranking of the surveyed systems according to his/her own priorities and based on the overall properties of the surveyed systems (Tables 3, 4, 5, 6, 7, and 8).

5.1 Detection of the Face and Its Features

Most of the currently existing systems for facial expression analysis assume that the presence of a face in the scene is ensured. However, in many instances, the systems do not utilize a camera setting that will ascertain the correctness of that assumption. Only two of the surveyed systems process images acquired by a mounted camera ([59], [62]) and only two systems deal with the automatic face detection in an arbitrary scene ([30], [24]). In addition, in many instances strong assumptions are made to make the problem of facial expression analysis more tractable (Table 6). Some common assumptions are: the images contain frontal facial view, the illumination is constant, the light source is fixed, the face has no facial hair or glasses, the subjects are young (i.e., without permanent wrinkles), and of the same ethnicity. Also, in most of the real-life situations, it cannot be assumed that the observed subject will remain immovable, as assumed by some methods. Thus, if a fixed camera acquires the images, the system should be capable of dealing with rigid head motions. Only three of the surveyed systems deal to some extent with rigid head motions ([2], [18], [30]). For the sake of universality, the system should be able to analyze facial expressions of any person independently of age, ethnicity, and outlook. Yet, only the method proposed by Essa and Pentland [24] deals with the facial images of faces with facial hair and/or eyeglasses.

For the researchers of automated vision-based facial expression analysis, this suggest investigating towards a robust detection of the face and its features despite the changes in viewing and lightning conditions and the distractions like glasses, facial hair or changes in hair style. Another interesting but yet not investigated ability of the human visual system is "filling in" missing parts of the observed face and "perceiving" a whole face even when a part of it is occluded (e.g., by hand).

5.2 Facial Expression Classification

In general, the existing expression analyzers perform a singular classification of the examined expression into one of the basic emotion categories proposed by Ekman and Friesen [20]. This approach to expression classification has two main limitations. First, "pure" emotional expressions are seldom elicited. Most of the time, people show blends of emotional expressions. Therefore, classification of an expression into a single emotion category isn't realistic. An automated facial expression analyzer should realize quantified classification into multiple emotion categories. Only two of the surveyed systems ([62], [106]) perform quantified facial expression classification into multiple basic emotion categories. Second, it is not at all certain that all facial expressions displayed on the face can be classified under the six basic emotion categories. So even if an expression analyzer performs a quantified expression classification into multiple basic emotion categories, it would probably not be capable of interpreting each and every encountered expression. A psychological discussion on the topic can be found in [33], [76], [77], and [26]. Some experimental proofs can be found in the studies of Asian researchers (e.g., [32], [106]), which reported that their Asian subjects have difficulties to express some of the basic expressions (e.g., disgust and fear). Defining interpretation categories into which any facial expression can be classified is one of the key challenges in the design of a realistic facial expression analyzer. The lack of psychological scrutiny on the topic makes the problem even harder. A way of dealing with this problem is to build a system that learns its own expertise and allows the user to define his/her own interpretation categories (e.g., see [40]).

If the system is to be used for behavioral science investigations of the face, it should perform expression recognition as applied to automated FACS encoding. In other words, it should accomplish facial action coding from input images and quantification of those codes [16], [17]. Four of the surveyed systems perform facial action coding from an input image or an image sequence (Table 7). Yet, none of these systems performs quantification of the facial action codes. This task is particularly difficult to accomplish for a number of reasons. First, there are merely five different AUs for which FACS provides an option to score intensity on a 3-level intensity scale (low, medium, and

high). Second, some facial actions such as blink, wink, and lips sucked into the mouth, are either encountered or not. It is not reasonable to talk about a blink having a "higher intensity" than another blink. In addition, each person has his/her own maximal intensity of displaying a particular facial action. Therefore, it should be aimed on design of a system that can start with a generic facial action classification and then adapt to a particular individual to perform a quantification of the accomplished coding for the facial actions for which measuring the activation intensity is "reasonable." Also, none of the surveyed systems can distinguish all 44 AUs defined in FACS. This remains a key challenge for the researchers of automated FACS encoding. Another appealing but still not investigated property of the human visual system is assigning a higher "priority" to the upper face features than to the lower face features since they play a more important role in facial expression interpretation [22].

6 Conclusion

Analysis of facial expressions is an intriguing problem which humans solve with quite an apparent ease. We have identified three different but related aspects of the problem: face detection, facial expression information extraction, and facial expression classification. Capability of the human visual system in solving these problems has been discussed. It should serve as a reference point for any automatic vision-based system attempting to achieve the same functionality. Among the problems, facial expression classification has been studied most, due to its utility in application domains of human behavior interpretation and HCI. Most of the surveyed systems, however, are based on frontal view images of faces without facial hair and glasses what is unrealistic to expect in these application domains. Also, all of the proposed approaches to automatic expression analysis perform only facial expression classification into the basic emotion categories defined by Ekman and Friesen [20]. Nevertheless, this is unrealistic since it is not at all certain that all facial expressions able to be displayed on the face can be classified under the six basic emotion categories. Furthermore, some of the surveyed methods have been tested only on the set of images used for training. We hesitate in belief that those systems are person-independent what, in turn, should be a basic property of a behavioral science research tool or of an advanced HCI. All the discussed problems are intriguing and none has been solved, in the general case. We expect that they would remain interesting to the researchers of automated vision-based facial expression analysis for some time.

Acknowledgments

The authors would like to thank J. Steffens, M.J. Lyons, and T. Otsuka for providing us their high quality images used in the paper, and Dr. Kevin Bowyer as well as the anonymous reviewers for helping to improve the paper. Also, we would like to thank the IEEE Press and the Academic Press for granting us permission to reprint the figures appearing in the paper.

References

[1] J.N. Bassili, "Facial Motion in the Perception of Faces and of Emotional Expression," *J. Experimental Psychology 4,* pp. 373-379, 1978.

[2] M.J. Black and Y. Yacoob, "Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion," *Int'l J. Computer Vision,* vol. 25, no. 1, pp. 23-48, 1997.

[3] M.J. Black and Y. Yacoob, "Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using Local Parametric Models of Image Motions," *Proc. Int'l Conf. Computer Vision,* pp. 374-381, 1995.

[4] E. Boyle, A.H. Anderson, and A. Newlands, "The Effects of Visibility on Dialogue in a Cooperative Problem Solving Task," *Language and Speech,* vol. 37, no. 1, pp. 1-20, 1994.

[5] V. Bruce, *Recognizing Faces.* Hove, East Sussex: Lawrence Erlbaum Assoc., 1986.

[6] V. Bruce, "What the Human Face Tells the Human Mind: Some Challenges for the Robot-Human Interface," *Proc. Int'l Workshop Robot and Human Comm.,* pp. 44-51, 1992.

[7] J. Buhmann, J. Lange, and C. von der Malsburg, "Distortion Invariant Object Recognition—Matching Hierarchically Labelled Graphs," *Proc. Int'l Joint Conf. Neural Networks,* pp. 155-159, 1989.

[8] F.W. Campbell, "How Much of the Information Falling on the Retina Reaches the Visual Cortex and How Much is Stored in the Memory?" *Seminar at the Pontificae Academiae Scientiarium Scripta Varia,* 1983.

[9] L.S. Chen, T.S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal Human Emotion/Expression Recognition," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 366-371, 1998.

[10] J.F. Cohn, A.J. Zlochower, J.J. Lien, and T. Kanade, "Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 396-401, 1998.

[11] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active Shape Models—Training and Application," *Computer Vision Image Understanding,* vol. 61, no. 1, pp. 38-59, 1995.

[12] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," *Proc. European Conf. Computer Vision,* vol. 2, pp. 484-498, 1998.

[13] G.W. Cottrell and J. Metcalfe, "EMPATH: Fface, Emotion, Gender Recognition Using Holons," *Advances in Neural Information Processing Systems 3,* R.P. Lippman, ed., pp. 564-571, 1991.

[14] D. DeCarlo, D. Metaxas, and M. Stone, "An Anthropometric Face Model Using Variational Techniques," *Proc. SIGGRAPH,* pp. 67-74, 1998.

[15] L.C. De Silva, T. Miyasato, and R. Nakatsu, "Facial Emotion Recognition Using Multimodal Information," *Proc. Information, Comm., and Signal Processing Conf.,* pp. 397-401, 1997.

[16] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 10, pp. 974-989, Oct. 1999.

[17] M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Measuring Facial Expressions by Computer Image Analysis," *Psychophysiology,* vol. 36, pp. 253-263, 1999.

[18] G.J. Edwards, T.F. Cootes, and C.J. Taylor, "Face Recognition Using Active Appearance Models," *Proc. European Conf. Computer Vision,* vol. 2, pp. 581-695, 1998.

[19] P. Eisert and B. Girod, "Analysing Facial Expressions for Virtual Conferencing," *IEEE Trans Computer Graphics and Applications,* vol. 18, no. 5, pp. 70-78, 1998.

[20] P. Ekman and W.V. Friesen, *Unmasking the Face.* New Jersey: Prentice Hall, 1975.

[21] P. Ekman and W.V. Friesen, *Facial Action Coding System (FACS): Manual.* Palo Alto: Consulting Psychologists Press, 1978.

[22] P. Ekman, *Emotion in the Human Face.* Cambridge Univ. Press, 1982.

[23] H.D. Ellis, "Process Underlying Face Recognition," *The Neuropsychology of Face Perception and Facial Expression,* R. Bruyer, ed. pp. 1-27, New Jersey: Lawrence Erlbaum Assoc., 1986.

[24] I. Essa and A. Pentland, "Coding, Analysis Interpretation, Recognition of Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 757-763, July 1997.

[25] A.J. Fridlund, P. Ekman, and H. Oster, "Facial Expressions of Emotion: Review Literature 1970-1983," *Nonverbal Behavior and Communication,* A.W. Siegman and S. Feldstein, eds., pp. 143-224. Hillsdale NJ: Lawrence Erlbaum Assoc., 1987.

[26] A.J. Fridlund, "Evolution and Facial Action in Reflex, Social Motive, and Paralanguage," *Biological Psychology,* vol. 32, pp. 3-100, 1991.

[27] D.J. Hand, *Discrimination and Classification.* John Wiley and Sons, 1981.

[28] F. Hara and H. Kobayashi, "State of the Art in Component Development for Interactive Communication with Humans," *Advanced Robotics,* vol. 11, no. 6, pp. 585-604, 1997.

[29] R.J. Holt, T.S. Huang, A.N. Netravali, and R.J. Qian, "Determining Articulated Motion from Perspective Views," *Pattern Recognition,* vol. 30, no. 9, pp. 1435-1,449, 1997.

[30] H. Hong, H. Neven, and C. von der Malsburg, "Online Facial Expression Recognition Based on Personalized Galleries," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 354-359, 1998.

[31] B. Horn and B. Schunck, "Determining Optical Flow," *Artificial Intelligence,* vol. 17, pp. 185-203, 1981.

[32] C.L. Huang and Y.M. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification," *J. Visual Comm. and Image Representation,* vol. 8, no. 3, pp. 278-290, 1997.

[33] C.E. Izard, *The Face of Emotion.* New York: Appleton-Century-Crofts, 1971.

[34] C.E. Izard, "Facial Expressions and the Regulation of Emotions," *J. Personality and Social Psychology,* vol. 58, no. 3, pp. 487-498, 1990.

[35] T. Johanstone, R. Banse, and K.S. Scherer, "Acoustic Profiles in Prototypical Vocal Expressions of Emotions," *Proc. Int'l Conf. Phonetic Science,* vol. 4, pp. 2-5, 1995.

[36] I. Kanter and H. Sompolinsky, "Associative Recall of Memory without Errors" *Physical Review,* vol. 35, no. 1, pp. 380-392, 1987.

[37] M. Kass, A. Witkin, and D. Terzopoulos, "Snake: Active Contour Model," *Proc. Int'l Conf. Computer Vision,* pp. 259-269, 1987.

[38] M. Kato, I. So, Y. Hishinuma, O. Nakamura, and T. Minami, "Description and Synthesis of Facial Expressions Based on Isodensity Maps," *Visual Computing,* T. Kunii, ed., pp. 39-56. Tokyo: Springer-Verlag, 1991.

[39] F. Kawakami, M. Okura, H. Yamada, H. Harashima, and S. Morishima, "3D Emotion Space for Interactive Communication," *Proc. Computer Science Conf.,* pp. 471-478, 1995.

[40] G.D. Kearney and S. McKenzie, "Machine Interpretation of Emotion: Design of Memory-Based Expert System for Interpreting Facial Expressions in Terms of Signaled Emotions (JANUS)," *Cognitive Science,* vol. 17, no. 4, pp. 589-622, 1993.

[41] S. Kimura and M. Yachida, "Facial Expression Recognition and Its Degree Estimation," *Proc. Computer Vision and Pattern Recognition,* pp. 295-300, 1997.

[42] H. Kobayashi and F. Hara, "Facial Interaction between Animated 3D Face Robot and Human Beings," *Proc. Int'l Conf. Systems, Man, Cybernetics,,* pp. 3,732-3,737, 1997.

[43] H. Kobayashi and F. Hara, "Recognition of Six Basic Facial Expressions and Their Strength by Neural Network," *Proc. Int'l Workshop Robot and Human Comm.,* pp. 381-386, 1992.

[44] H. Kobayashi and F. Hara, "Recognition of Mixed Facial Expressions by Neural Network," *Proc. Int'l Workshop Robot and Human Comm.,* pp. 387-391, 1992.

[45] K.M. Lam and H. Yan, "An Analytic-to-Holistic Approach for Face Recognition Based on a Single Frontal View," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 7, pp. 673-686, July 1998.

[46] H.K. Lee and J.H. Kim, "An HMM-Based Threshold Model Approach for Gesture Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 10, pp. 961-973, Oct. 1999.

[47] H. Li and P. Roivainen, "3D Motion Estimation in Model-Based Facial Image Coding," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 6, pp. 545-555, 1993.

[48] J.J. Lien, T. Kanade, J.F. Cohn, and C.C. Li, "Automated Facial Expression Recognition Based on FACS Action Units," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 390-395, 1998.

[49] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *Proc. Joint Conf. Artificial Intelligence,* pp. 674-680, 1981.

[50] M.J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 200-205, 1998.

[51] M.J. Lyons, J. Budynek, and S. Akamatsu, "Automatic Classification of Single Facial Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 12, pp. 1,357-1,362, 1999.

[52] K. Mase, "Recognition of Facial Expression from Optical Flow," *IEICE Trans.,* vol. E74, no. 10, pp. 3,474-3,483, 1991.

[53] K. Matsumura, Y. Nakamura, and K. Matsui, "Mathematical Representation and Image Generation of Human Faces by Metamorphosis," *Electronics and Comm. in Japan—3,* vol. 80, no. 1, pp. 36-46, 1997.

[54] K. Matsuno, C.W. Lee, and S. Tsuji, "Recognition of Facial Expression with Potential Net," *Proc. Asian Conf. Computer Vision,* pp. 504-507, 1993.

[55] A. Mehrabian, "Communication without Words," *Psychology Today,* vol. 2, no. 4, pp. 53-56, 1968.

[56] S. Morishima, F. Kawakami, H. Yamada, and H. Harashima, "A Modelling of Facial Expression and Emotion for Recognition and Synthesis," *Symbiosis of Human and Artifact,* Y. Anzai, K. Ogawa and H. Mori, eds., pp. 251-256, Amsterdam: Elsevier Science BV, 1995.

[57] Y. Moses, D. Reynard, and A. Blake, "Determining Facial Expressions in Real Time," *Proc. Int'l Conf. Automatic Face and GestureRecognition,* pp. 332-337, 1995.

[58] R. Nakatsu, "Toward the Creation of a New Medium for the Multimedia Era," *Proc. IEEE,* vol. 86, no. 5, pp. 825-836, 1998.

[59] T. Otsuka and J. Ohya, "Recognition of Facial Expressions Using HMM with Continuous Output Probabilities," *Proc. Int'l Workshop Robot and Human Comm.,* pp. 323-328, 1996.

[60] T. Otsuka and J. Ohya, "Spotting Segments Displaying Facial Expression from Image Sequences Using HMM," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 442-447, 1998.

[61] C. Padgett and G.W. Cottrell, "Representing Face Images for Emotion Classification," *Proc. Conf. Advances in Neural Information Processing Systems,* pp. 894-900, 1996.

[62] M. Pantic and L.J.M. Rothkrantz, "Expert System for Automatic Analysis of Facial Expression," *Image and Vision Computing J.,* vol. 18, no. 11, pp. 881-905, 2000.

[63] M. Pantic and L.J.M. Rothkrantz, "An Expert System for Multiple Emotional Classification of Facial Expressions" *Proc. Int'l Conf. Tools with Artificial Intelligence,* pp. 113-120, 1999.

[64] V.I. Pavlovic, R. Sharma, and T.S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: Review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 677-695, 1997.

[65] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. Computer Vision and Pattern Recognition,* pp. 84-91, 1994.

[66] V.A. Petrushin, "Emotion in Speech: Recognition and Application to Call Centers," *Proc. Conf. Artificial Neural Networks in Eng.,* 1999.

[67] R.W. Picard and E. Vyzas, "Offline and Online Recognition of Emotion Expression from Physiological Data," *Emotion-Based Agent Architectures Workshop Notes, Int'l Conf. Autonomous Agents,* pp. 135-142, 1999.

[68] T.S. Polzin and A.H. Waibel, "Detecting Emotions in Speech," *Proc. Conf. Cooperative Multimedia Comm.,* 1998.

[69] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C,* Cambridge Univ. Press, 1992.

[70] A. Rahardja, A. Sowmya, and W.H. Wilson, "A Neural Network Approach to Component versus Holistic Recognition of Facial Expressions in Images," *SPIE, Intelligent Robots and Computer Vision X: Algorithms and Techniques,* vol. 1,607, pp. 62-70, 1991.

[71] A. Ralescu and R. Hartani, "Some Issues in Fuzzy and Linguistic Modeling" *Proc. Conf. Fuzzy Systems,* pp. 1,903-1,910, 1995.

[72] M. Riedmiller and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm," *Proc. Int'l Conf. Neural Networks,* pp. 586-591, 1993.

[73] M. Rosenblum, Y. Yacoob, and L. Davis, "Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture," *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects,* pp. 43-49, 1994.

[74] H.A. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 1, pp. 23-38, Jan. 1998.

[75] *The Psychology of Facial Expression,* J.A. Russell and J.M. Fernandez-Dols, eds. Cambridge: Cambridge Univ. Press, 1997.

[76] J.A. Russell, "Is There Universal Recognition of Emotion from Facial Expression?" *Psychological Bulletin,* vol. 115, no. 1, pp. 102-141, 1994.

[77] P. Ekman, "Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique," *Psychological Bulletin,* vol. 115, no. 2, pp. 268-287, 1994.

[78] A. Samal, "Minimum Resolution for Human Face Detection and Identification," *SPIE Human Vision, Visual Processing, and Digital Display II,* vol. 1,453, pp. 81-89, 1991.

[79] A. Samal and P.A. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey," *Pattern Recognition,* vol. 25, no. 1, pp. 65-77, 1992.

[80] E. Simoncelli, "Distributed Representation and Analysis of Visual Motion," PhD thesis, Massachusetts Inst. of Technology, 1993.

[81] J. Steffens, E. Elagin, and H. Neven, "PersonSpotter—Fast and Robust System for Human Detection, Tracking, and Recognition," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 516-521, 1998.

[82] G.M. Stephenson, K. Ayling, D.R. Rutter, "The Role of Visual Communication in Social Exchange," *Britain J. Social Clinical Psychology,* vol. 15, pp. 113-120, 1976.

[83] K.K. Sung and T. Poggio, "Example-Based Learning for View-Based Human Face Detection" *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 1, pp. 39-51, Jan. 1998.

[84] A. Takeuchi and K. Nagao, "Communicative Facial Displays as a New Conversational Modality," *Proc. ACM INTERCHI,* pp. 187-193, 1993.

[85] J.C. Terrillon, M. David, and S. Akamatsu, "Automatic Detection of Human Faces in Natural Scene Images by Use of a Skin Color Model of Invariant Moments" *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 112-117, 1998.

[86] D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 6, pp. 569-579, June 1993.

[87] N.M. Thalmann and D. Thalmann, "600 Indexed References on Computer Animation," *J. Visualisation and Computer Animation,* vol. 3, pp. 147-174, 1992.

[88] N.M. Thalmann, P. Kalra, and M. Escher, "Face to Virtual Face," *Proc. IEEE,* vol. 86, no. 5, pp. 870-883, 1998.

[89] N.M. Thalmann, P. Kalra, and I.S. Pandzic, "Direct Face-to-Face Communication between Real and Virtual Humans" *Int'l J. Information Technology,* vol. 1, no. 2, pp. 145-157, 1995.

[90] N. Tosa and R. Nakatsu, "Life-Like Communication Agent—Emotion Sensing Character MIC and Feeling Session Character MUSE," *Proc. Conf. Multimedia Computing and Systems,* pp. 12-19, 1996.

[91] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[92] H. Ushida, T. Takagi, and T. Yamaguchi, "Recognition of Facial Expressions Using Conceptual Fuzzy Sets" *Proc. Conf. Fuzzy Systems,* vol. 1, pp. 594-599, 1993.

[93] P. Vanger, R. Honlinger, and H. Haken, "Applications of Synergetics in Decoding Facial Expression of Eemotion," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 24-29, 1995.

[94] J.M. Vincent, D.J. Myers, and R.A. Hutchinson, "Image Feature Location in Multi-Resolution Images Using a Hierarchy of Multi-Layer Preceptors," *Neural Networks for Speech, Vision, and Natural Language,* pp. 13-29, Chapman & Hall, 1992.

[95] M. Wang, Y. Iwai, and M. Yachida, "Expression Recognition from Time-Sequential Facial Images by Use of Expression Change Model," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 324-329, 1998.

[96] D.J. Williams and M. Shah, "A Fast Algorithm for Active Contours and Curvature Estimation," *Computer Vision and Image Processing: Image Understanding,* vol. 55, no. 1, pp. 14-26, 1992.

[97] A.D. Wilson and A.F. Bobick, "Parametric Hidden Markov Models for Gesture Recognition" *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 9, pp. 884-900, Sept. 1999.

[98] L. Wiskott, "Labelled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis," *Reihe Physik,* vol. 53, Frankfurt a.m. Main: Verlag Harri Deutsch, 1995.

[99] H. Wu, T. Yokoyama, D. Pramadihanto, and M. Yachida, "Face and Facial Feature Extraction from Color Image," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 345-350, 1996.

[100] Y. Yacoob and L. Davis, "Recognizing Facial Expressions by Spatio-Temporal Analysis," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 747-749, 1994.

[101] Y. Yacoob and L. Davis, "Computing Spatio-Temporal Representations of Human Faces," *Proc. Computer Vision and Pattern Recognition,* pp. 70-75, 1994.

[102] H. Yamada, "Visual Information for Categorizing Facial Expressions of Emotions," *Applied Cognitive Psychology,* vol. 7, pp. 257-270, 1993.

[103] J. Yang and A. Waibel, "A Real-Time Face Tracker," *Workshop Applications of Computer Vision,* pp. pp. 142-147, 1996.

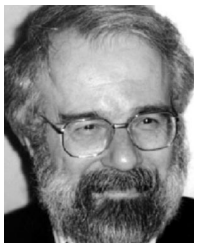
[104] M. Yoneyama, Y. Iwano, A. Ohtake, and K. Shirai, "Facial Expressions Recognition Using Discrete Hopfield Neural Networks," *Proc. Int'l Conf. Information Processing,* vol. 3, pp. 117-120, 1997.

[105] A.L. Yuille, D.S. Cohen, and P.W. Hallinan, "Feature Extraction from Faces Using Deformable Templates," *Proc. Computer Vision and Pattern Recognition,* pp. 104-109, 1989.

[106] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 454-459, 1998.

[107] J. Zhao and G. Kearney, "Classifying Facial Emotions by Backpropagation Neural Networks with Fuzzy Inputs," *Proc. Conf. Neural Information Processing,* vol. 1, pp. 454-457, 1996.

Maja Pantic received the MS degree in computer science (cum laude) from the Department of Computer Science, Faculty of Technical Mathematics and Computer Science at the Delft University of Technology, the Netherlands, in 1997. She is a PhD candidate in computer science at the Department of Computer Science, Faculty of Information Technology and Systems at the Delft University of Technology, the Netherlands. Her research interests are in the areas of multimedia multimodal man-machine interfaces and artificial intelligence including knowledge-based systems, distributive AI, machine learning, and applications of artificial intelligence in intelligent multimodal user interfaces. She is a student member of the IEEE..



Leon Rothkrantz received the MSc degree in mathematics from the University of Utrecht, the Netherlands, in 1971, the PhD degree in mathematics from the University of Amsterdam, the Netherlands, in 1980, and the MSc degree in psychology from the University of Leiden, the Netherlands, in 1990. As an associate professor, he joined the Knowledge Based Systems group of the Faculty of Technical Mathematics and Computer Science at the Delft University of Technology, the Netherlands, in 1992. The long range goal of Dr. Rothkrantz's research is design and development of human-psychology-based anthropomorphic multimodal multimedia fourth generation man-machine interface. His interests include applying computational technology to analysis of all aspects of human behavior.