Pattern Anal Applic (2008) 11:101–116
DOI 10.1007/s10044-007-0084-8

THEORETICAL ADVANCES

Recognising facial expressions in video sequences

José M. Buenaposada · Enrique Muñoz · Luis Baumela

Received: 7 January 2007 / Accepted: 10 July 2007 / Published online: 18 October 2007

Abstract We introduce a system that processes a sequence of images of a front-facing human face and recognises a set of facial expressions. We use an efficient appearance-based face tracker to locate the face in the image sequence and estimate the deformation of its non-rigid components. The tracker works in real time. It is robust to strong illumination changes and factors out changes in appearance caused by illumination from changes due to face deformation. We adopt a model-based approach for facial expression recognition. In our model, an image of a face is represented by a point in a deformation space. The variability of the classes of images associated with facial expressions is represented by a set of samples which model a low-dimensional manifold in the space of deformations. We introduce a probabilistic procedure based on a nearest-neighbour approach to combine the information provided by the incoming image sequence with the prior information stored in the expression manifold to compute a posterior probability associated with a facial expression. In the experiments conducted we show that this system is able to work in an unconstrained environment with strong changes in illumination and face location. It achieves an 89% recognition rate in a set of 333 sequences from the Cohn–Kanade database.

Keywords Facial expression recognition · Manifold of facial expressions · Nearest neighbour

1 Introduction

In recent years, industry and academia have shown growing interest in the development of computer vision systems that can locate human faces, track their motion and recognise their facial expressions. This interest is based on the fact that this technology is a key component in the development of advanced human–computer interaction systems.

One of the challenges of computer science is to make computers that interact with humans in a natural way, as humans interact with each other. Spoken language is possibly one of the most natural ways of interaction, but, unfortunately, it is ambiguous. This is why human interaction is based on two channels [19]. The first one, based on language, transmits explicit information. The other, based on gestures and facial expressions, transmits implicit information on how to interpret what is transmitted through the explicit channel. The context and the information provided by the implicit channel are extremely important for computers to get a full understanding of what is actually transmitted in a conversation. For example, the sentence "that will do" uttered by a customer could be interpreted by an Internet sales software agent as a request for more information, if the customer has a facial expression conveying inquiry or concern, or as a confirmation of a purchase if the costumer nods. The introduction of emotive icons in email messages is also a recent example of the necessity of implicit information in an explicit message.

J. M. Buenaposada
ESCET, Universidad Rey Juan Carlos,
C/Tulipán s/n, 28933 Móstoles, Spain

E. Muñoz · L. Baumela (✉)
Facultad Informática, Universidad Politécnica de Madrid,
Campus Montegancedo s/n, 28660 Boadilla del Monte, Spain
e-mail: lbaumela@fi.upm.es
URL: http://www.dia.fi.upm.es/∼pcr

Present Address:
E. Muñoz
Facultad Informática, Universidad Complutense de Madrid,
Ciudad Universitaria s/n, 28040 Madrid, Spain

An enormous body of research and important achievements have been made over the past 40 years within the natural language and speech recognition research communities on developing computer systems capable of decoding the explicit channel [39]. On the other hand, the decodification of the implicit channel has not received much attention until more recently [19, 47]. It is a challenging problem, since it is associated with the understanding of a person's intentions and emotions and requires close collaboration between computer vision, pattern and speech recognition, psychology and linguistics. Some systems use physiological signals [49] as raw input for emotion classification although most systems are based on audiovisual information [1] because of the non-invasive nature of this signal. In this paper, we describe a system that boosts several state-of-the-art aspects of decoding the implicit channel from a computer vision perspective.

For over 30 years Paul Ekman and his colleagues have studied human facial expressions and their relation to emotions [23]. They suggest that there is evidence to support the existence of six primary emotions, which are universal across cultures and human ethnicities [24]. Each emotion possesses a distinctive prototypic facial expression. These basic emotions are joy (jo), surprise (su), anger (an), sadness (sa), fear (fe) and disgust (di). Recognising all or a subset of these prototypic facial expressions from images has been a topic of research in computer vision and pattern recognition for the last decade [5, 11, 16, 17, 25, 44, 52, 63, 65].

In this paper, we describe a system that tracks the rigid motion of a front-facing human face in real time, while estimating the deformation of its non-rigid elements. The descriptors representing the non-rigid deformation of the face are used to estimate the facial expression. Our work focuses on both, the use of an efficient non-rigid face tracker, robust to strong changes in the scene illumination, and the construction of a classifier to probabilistically recognise prototypic facial expressions in video sequences.

Tracking a human face is a challenging problem because the face is a deformable low-textured object and because its visual appearance changes dramatically from one person to another and in the presence of occlusions, changes in illumination or pose. In this paper, we adopt a model-based procedure for tracking. In our approach the appearance of a face is represented by the addition of two approximately independent linear subspaces. The first subspace models the deformations of the face caused by facial expressions. The second represents the variations in facial appearance caused by changes in the scene illumination. The tracker presented here is simple, efficient, robust and user dependent. All the information to be provided to particularise this tracker for a new user is a front-facing picture of the user wearing more or less a neutral expression.

We also adopt a model-based approach for facial expression recognition. By tracking a set of 333 image sequences from 92 different users from the Cohn–Kanade database [33], we build a user-and-illumination-independent global representation of all facial expressions. In this model, an image of a face is considered as a point in an n-dimensional space of deformations (n is the number of face tracker parameters). The variability of the classes of images associated with the prototypic facial expressions is represented by a set of samples that model a low-dimensional manifold embedded in the n-dimensional space of deformations. Pictures representing similar expressions are mapped to nearby points on the manifold. An image sequence becomes a path in the space of deformations. In order to recognise the facial expressions in the sequence, we introduce a probabilistic procedure based on a nearest-neighbour approach to combine the information provided by the image sequence with the prior information represented in the manifold. For each prototypic expression, we estimate a posterior probability, given the images in the sequence and the manifold of the expression. At a given time instant, the most likely expression is given by the maximum of these posterior probabilities.

In Sect. 6, we show that this system achieves recognition results in the Cohn–Kanade image database similar to the best state-of-the-art systems. Moreover, our system is able to work in an unconstrained environment, with strong variations in illumination and fast and large in-plane and small out-of-plane rigid head motion.

The rest of this paper is organised as follows. In the following section, we present related work. In Sect. 3 we describe the face detection and tracking algorithm. The manifold of facial expressions and the expression recognition procedures are described in Sects. 4 and 5 respectively. In Sect. 6 we describe some of the experiments that we have conducted on this system, and, finally, in Sect. 7 we draw conclusions.

2 Related work

The problem of facial expression recognition can be divided into three subproblems: face detection, discriminative information extraction and expression classification. Face detection aims at locating faces in complex scenes and cluttered backgrounds. Video-based facial expression recognition techniques use face trackers to locate the face in each image in the sequence. In this case, face detection algorithms are used to start-up the tracking procedure or to recover the tracker from a complete loss. Once the position of the face in an image has been estimated, it is analysed to extract discriminative information that will subsequently be used to classify the facial expression. Different facial

expression recognition algorithms have been introduced in the literature depending on the discriminative information extracted from the image and the classification procedure used (see [26, 45] for a comprehensive survey). Here, we will review the algorithms that are most closely related to our work. We will not address the problem of face detection [48, 54, 62], since it has traditionally been treated as a separate problem from facial expression recognition.

Facial expressions are generated by contractions of facial muscles that deform facial elements such as eye-lids, eyebrows, nose, lips and skin texture. Feature-based approaches to facial deformation estimation extract discriminative information about the deformation of these facial elements from a discrete set of locations in the face. Initial feature-based approaches used make-up emphasised contours of eyebrows and lips [6], the corners of mouth, eyes and nostrils [28], colour markers [42] or the geometrical distribution of a set of fiducial points on the face [69]. Other approaches used a set of geometrical features on lips, eyebrows, cheek and furrow [59] or applied knowledge-based systems to reason about such features [46]. In a more recent paper, face line edges are used as features in a static face [27]. Many alternative attempts have focused on optical flow analysis [25, 37] estimated in textured areas [52, 65] or as local parametric models of motion [11].

Model-based approaches establish a set of modes of face deformation based on anatomically [58] or statistically [34] motivated data. An evolution of the statistical approach is the 2D [18] and 3D [12] shape plus texture linear models. The deformation of these models is estimated from the motion of the face's contours [58], from optical flow [20] or from the sum of squared differences of image grey values [38, 41, 50]. More recently 3D primitive surface description features have also been used for expression recognition from range data [63].

Other methods estimate facial motion or deformation from the analysis of pixel grey level values on face areas. This is the case of Gabor filters, which are robust to illumination changes and detect face edges on multiple scales and with different orientations [4, 5, 21, 36, 51, 69], the local binary patterns (LBP) [56] and volumetric local binary patterns [70] and also of the eigenface approaches [61].

Feature-based approaches only estimate the motion of a discrete set of textured regions. Unless a dense set of artificial markers is used, they provide sparse information about the deformation of the face. This information may not be adequate for modelling important components of an expression, such as wrinkles and dimpling. In [69], Gabor filters were favourably compared against discrete geometrical models and can be considered among the best discriminative procedures [5], but their computation is both time and memory intensive. Optical flow and eigenface techniques provide rich and dense information about facial motion, but are easily disturbed by lighting changes, registration inaccuracy and motion discontinuities [67]. Shape plus texture models can be fitted in real time to a deforming face [38] and may factor out variations in illumination. Their major drawback is that they are difficult to build [3, 12]. In this paper, we introduce a linear face model that models changes in an image's grey values caused by facial deformation and illumination. It can be efficiently fitted to a target image in real time and can be automatically trained (see Sect. 3).

The discriminative information obtained by the above techniques is fed into a classification algorithm to recognise the facial expression. Two groups of classification techniques have been used depending on whether the discriminative information was extracted from a single static image or from a sequence of images. Neural nets are possibly the most popular classification procedure among the static approaches [59, 69]. Other procedures used are Tree augmented Naïve Bayes [17] and more recently AdaBoost together with support vector machines [5] and linear discriminant analysis [63]. Hidden Markov models are the most common approach among the procedures based on the analysis of an image sequence [17, 35, 44, 66]. They have been extensively used because of their ability to deal with time-dependent parameters and to provide time scale invariance. Radial basis function neural nets with recurrent input [52] and more recently Bayesian networks [60, 68] have also been used as an alternative for modelling temporal information.

The common limitation of the static approaches is that they do not capture the dynamic information in the facial expression. This is a key factor revealing information about the subject's emotional state [7]. An alternative dynamic approach consists of mapping facial expression images to low-dimensional manifolds associated with each primitive expression. The expression manifold acts as a prior probability distribution on the appearance of a facial expression. A statistical procedure is used to combine the prior information with the input image sequence to get a posterior probability associated with each primary facial expression. With this approach, a target sequence may be assigned to the facial expression with maximum posterior probability or, alternatively, it may be described as a probabilistic blending of the primary expressions, opening up the possibility of recognising mixed expressions [16, 56]. In this paper, we take this approach and introduce a procedure to build the expression manifold using the parameters produced by our illumination-independent face tracker. We also introduce a statistical approach for estimating the posterior probability of each expression. Our solution differs from previous related approaches [16, 56] in various ways: (a) our manifolds are user independent, while those introduced in [16] depended on the user's

identity; (b) we use the parameters of an illumination-independent linear model as discriminative information for expression classification, whereas active wavelet networks [31] in [16] and LBP features [43] in [56] are respectively used; (c) we use a procedure for estimating the posterior probabilities different from those in [16] and in [56].

3 Face tracking and feature extraction

The system presented in this paper is able to robustly track a human face and recognise the facial expressions in an unconstrained environment with sharp illumination changes. To achieve this aim, we use a robust tracking architecture that co-ordinates three trackers (see Fig. 1). It is organised in three levels of increasing complexity. The execution policy is very simple: when a tracker performs satisfactorily it tries to transfer control to a higher level tracker; whenever a tracker detects a target loss, it transfers control to a lower level tracker. At the lowest level of the hierarchy we have a face detector, which could be based on the popular haar-like features [62] or on the simpler and less robust colour features [13]. At mid-level we use a template-based rigid face tracker [14]. This face tracker determines whether the skin-coloured blob detected by the lowest level tracker is a front-facing face and provides the start-up information for the higher level tracker. At the highest level we have a subspace-based tracker, which we describe in this section. A preliminary version of this tracker appeared in [15].

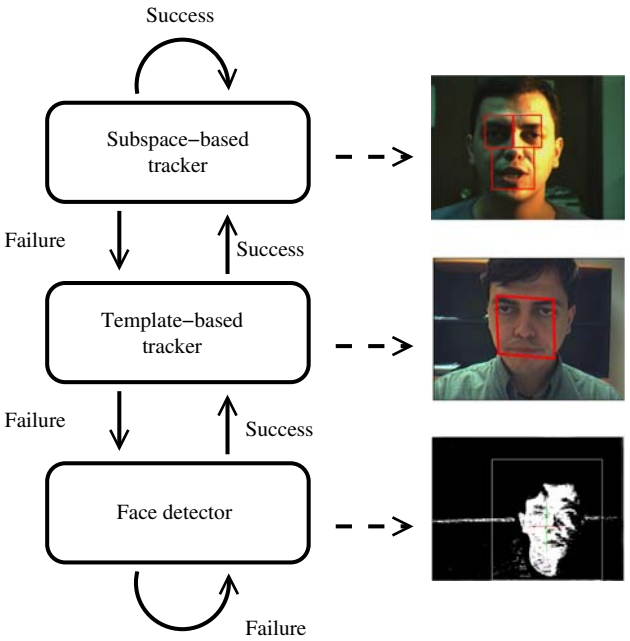


Fig. 1 Tracking architecture

3.1 A linear and illumination-independent face model

Here we introduce a subspace-based model representing the variations in the appearance of a face caused by changes in the facial expressions and the illumination of the scene.

Let $\mathtt{I}(\mathbf{x},t)$ be the image acquired at time t, where $\mathbf{x}$ is a vector representing the co-ordinates of a point in the image, and let $\mathtt{I}(\mathbf{x},t)$ be a vector storing the brightness values of $\mathtt{I}(\mathbf{x},t)$. Let us assume that the target moves rigidly (with no deformation) between time instants t_0 and t, and that this motion can be described by the motion model $f(\mathbf{x},\boldsymbol{\mu})$, $\boldsymbol{\mu}$ being the vector of rigid motion parameters. If there are no changes in the target appearance caused by the scene illumination, the brightness constancy equation $\mathtt{I}(f(\mathbf{x},\boldsymbol{\mu}_t),t) = \mathtt{I}(\mathbf{x},t_0)$ holds. If the face is now allowed to deform non-rigidly, then we may write a new brightness constancy equation $\mathtt{I}(f(\mathbf{x},\boldsymbol{\mu}_t),t) = \bar{\mathtt{I}}(\mathbf{x}) + [\mathtt{B}_d\mathbf{c}_{d,t}](\mathbf{x})$, where the non-rigid deformations have been modelled by a linear subspace with basis $\mathtt{B}_d$, mean value $\bar{\mathtt{I}}(\mathbf{x})$ and linear deformation parameters $\mathbf{c}_{d,t}$. We denote the value of $\mathtt{B}_d\mathbf{c}_{d,t}$ for the pixel with position $\mathbf{x}$ by $[\mathtt{B}_d\mathbf{c}_{d,t}](\mathbf{x})$. Finally, for a given rigid motion $\boldsymbol{\mu}_t$ and deformation $\mathbf{c}_{d,t}$, we could also model the illumination of the face by including a new subspace with basis $\mathtt{B}_i$ and linear illumination parameters $\mathbf{c}_i$, which represents all the possible illuminations of the mean face $\bar{\mathtt{I}}(\mathbf{x})$. So, the final brightness constancy equation is

$$\begin{aligned}\mathtt{I}(f(\mathbf{x},\boldsymbol{\mu}_t),t) &= \bar{\mathtt{I}}(\mathbf{x}) + [\mathtt{B}_i\mathbf{c}_{i,t}](\mathbf{x}) + [\mathtt{B}_d\mathbf{c}_{d,t}](\mathbf{x}) \\ &= \bar{\mathtt{I}}(\mathbf{x}) + [\mathtt{B}\mathbf{c}_t](\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{F},\end{aligned} \tag{1}$$

where $\mathtt{B} = [\mathtt{B}_i|\mathtt{B}_d], \mathbf{c}_t^\top = (\mathbf{c}_{i,t}^\top, \mathbf{c}_{d,t}^\top)^\top, k = \dim(\mathbf{c}_t)$, and $\mathcal{F}$ represents the set of pixels of the face used for tracking. Vectors $\mathbf{c}_i$ and $\mathbf{c}_d$ are respectively the illumination and the deformation appearance parameters. The assumption that illumination and deformation subspaces are independent will simplify the training of the model. Instead of having to use image sequences in which all combinations of illuminations and facial expressions are present, we will only have to process two image sequences: one with one facial expression and all illuminations and another with one illumination and all facial expressions. A related result for a rigid face moving in 3D space has been introduced recently [64].

To validate the above model we run the following experiment. First we trained the tracker according to the procedure described later in this section. Then we manually selected the parameters of two facial expressions and two illuminations, and generated a set of intermediate illuminations and expressions by uniformly sampling the parameter space between those locations. We repeated this process three times. The results are shown in Fig. 2. In

spite of the model's linearity, it correctly generates the appearance of the faces.

3.2 Efficiently tracking the face

Tracking a face consists of estimating, for each image in the sequence, the values of the motion, $\boldsymbol{\mu}$, and appearance, $\mathtt{c}$, parameters which minimise the error function

$$E(\boldsymbol{\mu}, \mathbf{c}) = \|\mathtt{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t) - \bar{\mathtt{I}} - [\mathtt{B}\mathbf{c}_t](\mathbf{x})\|^2. \quad (2)$$

To make the previous minimisation robust to occlusions, the quadratic error norm can be replaced by a robust one (e.g. see [10, 29]). The objective of the robust norm is to limit the bias introduced in the minimisation by those pixels for which $|\mathtt{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t) - \bar{\mathtt{I}} - [\mathtt{B}\mathbf{c}_t](\mathbf{x})|$ has an unusually high value.

In general, it can be hard to minimise (2) as it defines a non-convex cost function. Black and Jepson [10] presented an iterative solution using a gradient descent procedure and a robust metric with increasing resolution levels. Their algorithm is not suitable for real-time performance, since the Jacobian of each incoming image has to be computed once on every frame for each level in the multi-resolution pyramid. Similar problems have been solved efficiently using Gauss–Newton minimisation [29, 38]. Hager and Belhumeur [29] introduced an efficient procedure for minimising (2) in the context of invariance to illumination changes by assuming $\nabla_{\mathbf{x}}[\mathtt{B}\mathbf{c}](\mathbf{x}) \approx 0$. This assumption is a valid approximation when modelling the illumination of a rigid head, but it cannot be reliably used for tracking faces whose appearance changes due to causes other than illumination. Here, we introduce an efficient procedure for minimising (2) without such a restriction.

To make Gauss–Newton iterations, $\mathtt{I}$ is expanded as a Taylor series at $(\boldsymbol{\mu}_t, \mathbf{c}_t, t)$, producing a new error function

$$E(\delta\boldsymbol{\mu}, \delta\mathbf{c}) = \|\mathtt{M}\delta\boldsymbol{\mu} + \mathtt{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t + \delta t) - \bar{\mathtt{I}} - \mathtt{B}(\mathbf{c}_t + \delta\mathbf{c})\|^2, \quad (3)$$

where $\mathtt{M} = \left[\frac{\partial \mathtt{I}(f(\mathbf{x},\boldsymbol{\mu}),t)}{\partial \boldsymbol{\mu}}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_t}\right]$ is the $N \times n (n = \dim(\boldsymbol{\mu}))$ Jacobian matrix of $\mathtt{I}$.

3.2.1 Jacobian matrix factorisation

One of the obstacles for minimising (3) online, during tracking, is the computational cost of estimating $\mathtt{M}$ for each frame. In this section, we will show that $\mathtt{M}$ can be factored into the product of two matrices, $\mathtt{M}_0\Sigma(\boldsymbol{\mu}, \mathbf{c})$, where $\mathtt{M}_0$ is a constant matrix, which can be computed off-line.

Each row $m_i(\boldsymbol{\mu}_t, \mathbf{c}_t)$ of $\mathtt{M}(\boldsymbol{\mu}_t, \mathbf{c}_t)$ can be written as the product

$$m_i(\boldsymbol{\mu}_t, \mathbf{c}_t) = \nabla_{\mathtt{f}}\mathtt{I}(f(\mathbf{x}_i, \boldsymbol{\mu}_t), t)^{\top} f_{\boldsymbol{\mu}}(\mathbf{x}_i, \boldsymbol{\mu}_t), \quad (4)$$

where

$$\nabla_{\mathtt{f}}\mathtt{I}(f(\mathbf{x}_i, \boldsymbol{\mu}_t), t)^{\top} = \left[\frac{\partial \mathtt{I}(\mathbf{y}, t)}{\partial \mathbf{y}}\Big|_{\mathbf{y}=f(\mathbf{x}_i,\boldsymbol{\mu}_t)}\right]$$

and

$$f_{\boldsymbol{\mu}}(\mathbf{x}_i, \boldsymbol{\mu}_t) = \left[\frac{\partial f(\mathbf{x}_i, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_t}\right].$$

Taking derivatives w.r.t. $\mathbf{x}$ on both sides of (1) we get

$$\nabla_{\mathtt{f}}\mathtt{I}(f(\mathbf{x}_i, \boldsymbol{\mu}_t), t)^{\top} f_{\mathbf{x}}(\mathbf{x}_i, \boldsymbol{\mu}_t) = \nabla_{\mathbf{x}}\bar{\mathtt{I}}(\mathbf{x}) + \nabla_{\mathbf{x}}[\mathtt{B}\mathbf{c}_t](\mathbf{x}), \quad (5)$$

where $f_{\mathbf{x}}(\mathbf{x}_i, \boldsymbol{\mu}_t) = \left[\frac{\partial f(\mathbf{x},\boldsymbol{\mu}_t)}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}_i}\right]$ and $\nabla_{\mathbf{x}}$ denotes the image gradient. Finally, from (4) and (5) we get a new expression for $\mathtt{M}$,

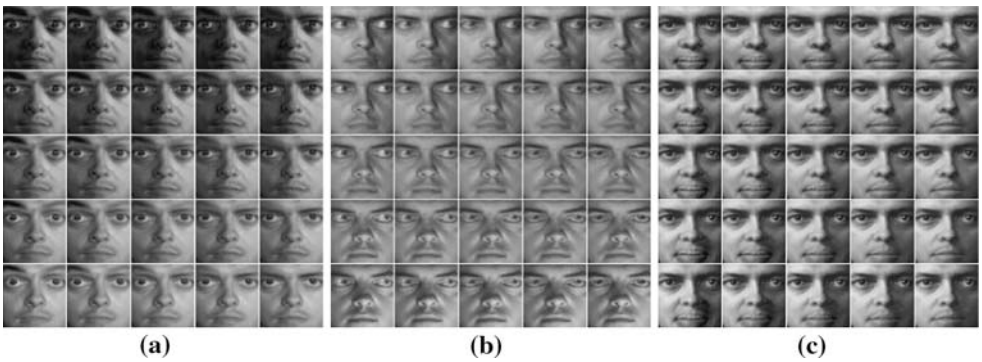


Fig. 2 Images generated using our appearance model. **a** From *left* to *right* images generated by lowering eyebrows, and from *top* to *down* images generated varying illumination; **b** rolling eyes with a different illumination; **c** closing mouth, again under different illumination

$$\mathtt{M}(\boldsymbol{\mu},\mathbf{c}) = \begin{bmatrix} \mathtt{B}_\nabla(\mathbf{x}_1)\mathtt{C}f_{\mathbf{x}}(\mathbf{x}_1,\boldsymbol{\mu})^{-1}f_{\boldsymbol{\mu}}(\mathbf{x}_1,\boldsymbol{\mu}) \\ \vdots \\ \mathtt{B}_\nabla(\mathbf{x}_N)\mathtt{C}f_{\mathbf{x}}(\mathbf{x}_N,\boldsymbol{\mu})^{-1}f_{\boldsymbol{\mu}}(\mathbf{x}_N,\boldsymbol{\mu}) \end{bmatrix}, \tag{6}$$

where $\mathtt{B}_\nabla$ is the gradient of the subspace basis vector and C is a matrix storing $\mathbf{c}$. Therefore, $\mathtt{M}$ can be expressed in terms of the gradient of the subspace basis vectors, $\mathtt{B}_\nabla$, which are constant, and the motion and appearance parameters $(\boldsymbol{\mu},\mathbf{c})$, which vary over time. If we choose a motion model f such that $\mathtt{C}f_{\mathbf{x}}(\mathbf{x}_i,\boldsymbol{\mu})^{-1}f_{\boldsymbol{\mu}}(\mathbf{x}_i,\boldsymbol{\mu}) = \Gamma(\mathbf{x}_i)\Sigma(\boldsymbol{\mu},\mathbf{c})$, then $\mathtt{M}$ can be factored into

$$\mathtt{M}(\boldsymbol{\mu},\mathbf{c}) = \begin{bmatrix} \mathtt{B}_\nabla(\mathbf{x}_1)\Gamma(\mathbf{x}_1) \\ \vdots \\ \mathtt{B}_\nabla(\mathbf{x}_N)\Gamma(\mathbf{x}_N) \end{bmatrix} \Gamma(\boldsymbol{\mu},\mathbf{c}) = \mathtt{M}_0\Sigma(\boldsymbol{\mu},\mathbf{c}), \tag{7}$$

where $\mathtt{M}_0$ is constant matrix and Σ depends on $\mathbf{c}$ and $\boldsymbol{\mu}$.

3.2.2 *Minimising* $E(\boldsymbol{\mu},\mathbf{c})$

The minimum of (3) can be estimated by least squares

$$\begin{bmatrix} \delta\boldsymbol{\mu} \\ \delta\mathbf{c} \end{bmatrix} = -(\mathtt{M}_J^\top\mathtt{M}_J)^{-1}\mathtt{M}_J\mathcal{E},$$

where $\mathtt{M}_J = (\mathtt{M}|-\mathtt{B})$ and $\mathcal{E} = \mathtt{I}(f(\mathbf{x},\boldsymbol{\mu}_t),t+\delta t) - \bar{\mathtt{I}} - \mathtt{B}\mathbf{c}_t$. Then, the change of rigid parameters may be estimated as $\delta\boldsymbol{\mu} = -(\mathtt{M}^\top\mathtt{N}_B\mathtt{M})^{-1}\mathtt{M}^\top\mathtt{N}_B\mathcal{E}$ and that of non-rigid parameters as $\delta\mathbf{c} = (\mathtt{B}^\top\mathtt{N}_M\mathtt{B})^{-1}\mathtt{B}^\top\mathtt{N}_M\mathcal{E}$, where $\mathtt{N}_B = \mathtt{I} - \mathtt{B}(\mathtt{B}^\top\mathtt{B})^{-1}\mathtt{B}^\top$ and $\mathtt{N}_M = \mathtt{I} - \mathtt{M}(\mathtt{M}^\top\mathtt{M})^{-1}\mathtt{M}^\top$. Since $\mathtt{N}_B$ is a constant matrix, we get an efficient solution for $\delta\boldsymbol{\mu}$ factoring $\mathtt{M}$ according to (7)

$$\delta\boldsymbol{\mu} = -(\Sigma^\top\Lambda_{M1}\Sigma)^{-1}\Sigma^\top\Lambda_{M2}\mathcal{E}, \tag{8}$$

where $\Lambda_{M1} = \mathtt{M}_0^\top\mathtt{N}_B\mathtt{M}_0$ and $\Lambda_{M2} = \mathtt{M}_0^\top\mathtt{N}_B$ are constant and can be precomputed off-line. A similar solution for $\delta\mathbf{c}$ would not be efficient, since $\mathtt{N}_M$ depends on $(\boldsymbol{\mu},\mathbf{c})$ and would have to be recomputed for each frame in the sequence. Nevertheless, an efficient solution can be obtained from (3) by least squares, considering that $\delta\boldsymbol{\mu}$ is known:

$$\delta\mathbf{c} = \Lambda_B[\mathtt{M}\delta\boldsymbol{\mu} + \mathcal{E}], \tag{9}$$

where $\Lambda_B = (\mathtt{B}^\top\mathtt{B})^{-1}\mathtt{B}^\top$ is also constant and can be precomputed off-line.

At first glance, this result may appear to be similar to the one presented in [38], Sect. 4.1, and in [29]. There are nevertheless three major differences: (a) here model parameters are additively updated, whereas in [38] the update procedure is compositional; (b) here subspace appearance parameters are incrementally estimated and additively updated $(\mathbf{c}_{t+1} = \delta\mathbf{c} + \mathbf{c}_t)$ and, consequently, $\mathcal{E}$ includes a $-\mathtt{B}\mathbf{c}_t$ term, whereas in either [38] or [29] there is no such term; (c) here the derivatives of the subspace basis are part of the Jacobian, whereas in [38] and in [29] they are not. As described in [29], this implies that assumption $\nabla_{\mathtt{x}}[\mathtt{B}\mathbf{c}](\mathtt{x}) \approx 0$. This assumption is approximately true for a rigid face, but not for a face whose appearance changes.

3.3 Subspace model building

One of the advantages of the appearance model introduced here is that the deformation and illumination subspaces are independent. A consequence of this property is that they can be independently trained. This allows us to simplify the training process. We do not need image sequences with all facial expressions under all possible illumination conditions. Now, each subspace is trained with one video sequence. For the illumination subspace we use a sequence in which a light orbits in front of the target face wearing a neutral expression. For the deformation subspace we use a sequence captured with a non-saturating frontal illumination in which the target face wears different facial expressions. The face is located and aligned in the first frame of both sequences. Then, with a procedure similar to the one described in [32], both sequences are independently tracked and both linear subspace models independently built (see Fig. 3).

Fig. 3 Some images used to build the deformation (*top row*) and illumination (*bottom row*) subspaces

3.4 The subspace-based tracking algorithm

In the implementation of our algorithm we use a RTS (rotation, translation and scale) motion model, so $\boldsymbol{\mu} = (\theta, t_u, t_v, s)$, and $f(\mathbf{x}, \boldsymbol{\mu}) = s\mathtt{R}(\theta)\mathbf{x} + \mathbf{t}$, where $\mathbf{x} = (u, v)^\top, \mathbf{t} = (t_u, t_v)^\top$ and $\mathtt{R}(\theta)$ is a 2D rotation matrix. In this case the factorisation in (7) results in

$$\Gamma(\mathbf{x}_i) = \left[\mathtt{I}_{2l\times 2l}, \begin{bmatrix} -v_i\mathtt{I}_{l\times l} & u_i\mathtt{I}_{l\times l} \\ u_i\mathtt{I}_{l\times l} & v_i\mathtt{I}_{l\times l}\end{bmatrix}\right],$$

$$\Sigma(\mathbf{c}, \boldsymbol{\mu}) = \begin{bmatrix} \mathtt{C}\frac{1}{s}\mathtt{R}(-\theta) & 0 \\ 0 & \mathtt{C}\begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{s}\end{bmatrix}\end{bmatrix},$$

where $\mathtt{I}_{d\times d}$ is the $d \times d$ identity matrix, C is a matrix storing $\mathbf{c}$ and $l = k + 1$, k being the dimension of the linear subspace. For this model $\mathtt{M}_0$ and Σ have dimensions $N \times 4l$ and $4l \times 4$, respectively.

The final factored modular tracking algorithm is shown in Algorithm 1.

Algorithm 1 Subspace tracking algorithm

Off-line:

Compute and store $\mathtt{M}_0$ using $\mathtt{B}$.
Compute and store $\Lambda_{M2} = \mathtt{M}_0^\top \mathtt{N}_B$.
Compute and store $\Lambda_{M1} = \Lambda_{M2}\mathtt{M}_0$.
Compute and store $\Lambda_B = (\mathtt{B}^\top\mathtt{B})^{-1}\mathtt{B}^\top$.

Online (one iteration):

Warp $\mathtt{I}(z, t + \delta t)$ to $\mathtt{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), \mathbf{t} + \delta\mathbf{t})$.
Compute $\mathcal{E} = [\mathtt{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t + \delta t) - \bar{\mathtt{I}} - \mathtt{B}\mathbf{c}_t]$.
Compute $\Sigma(\boldsymbol{\mu}_t, \mathbf{c}_t)$.
Compute $\mathtt{H} = \Sigma(\boldsymbol{\mu}_t, \mathbf{c}_t)^\top \Lambda_{M1}\Sigma(\boldsymbol{\mu}_t, \mathbf{c}_t)$.
Compute $\delta\boldsymbol{\mu} = -\mathtt{H}^{-1}\Sigma(\boldsymbol{\mu}_t, \mathbf{c}_t)^\top \Lambda_{M2}\mathcal{E}$.
Update $\boldsymbol{\mu}_{t+\delta t} = \boldsymbol{\mu}_t + \delta\boldsymbol{\mu}$.
Compute $\delta\mathbf{c}_{t+\delta t} = \Lambda_B[\mathtt{M}_0\Sigma(\boldsymbol{\mu}_t, \mathbf{c}_t)\delta\boldsymbol{\mu} + \mathcal{E}]$.
Update $\mathbf{c}_{t+\delta t} = \mathbf{c}_t + \delta\mathbf{c}_{t+\delta t}$.

4 The manifold of facial expressions

The classification procedure used for facial expression recognition is based on a user-and-illumination-independent facial expression model. This model is built by tracking a set of sequences from the Cohn–Kanade database [33]. This data set consists of 485 image sequences of 97 university students ranging in age from 18 to 30 years; 65% were female, 15% were African-American and 3% were Asian or Latino. Subjects began each display with a neutral face and ended it at the expression apex. The last image in each sequence is labelled with the FACS Action Units (AUs) [59] that describe the expression. We have manually translated these AUs into one of the six prototypic expressions. To construct our manifold, we selected 333 sequences of 92 subjects for which the prototypic expression could be clearly identified.

We used the tracker introduced in Sect. 3 to process the sequences from the database. The basis for the deformation and illumination subspaces of the tracker was obtained with the procedure and the training data described in Sect. 3.3. Although, as described in Sect. 3, our tracker was conceived to be dependent on the identity of the subject in the training sequence, we actually achieve a reasonable level of independence just by switching the average image, $\bar{\mathtt{I}}$, in (1) for an illumination-compensated picture of the new target subject wearing an approximately neutral expression. Let $\mathtt{I}_s$ be the image of the new subject and $\mathbf{c}_{i,s} = \mathtt{B}_i^\top(\mathtt{I}_s - \bar{\mathtt{I}})$ be the coefficients of the illumination of his or her face, then $\bar{\mathtt{I}}_s = \mathtt{I}_s - \mathtt{B}_i\mathbf{c}_{i,s}$ is the new average image. The intuition behind this is that, since $\bar{\mathtt{I}}$ is very similar to a picture of the subject (see Fig. 4), most of the information in the face model related to the subject's identity is stored in the mean face, whereas the information related to the facial expression is stored in the deformation subspace parameters $\mathbf{c}_d$. So, by just switching the mean image, we have a model for the new subject. We can use this new model to generate a picture of the new subject wearing the expression represented by parameters $\mathbf{c}_d$ (see Fig. 5). Although the results are not visually perfect, we will see in the experiments conducted in Sect. 6 that this new model is good enough to accurately track the subject and identify his or her facial expressions. Finally, to cancel other sources of appearance variation that are not directly related to facial expressions, we eliminate the eyes and the four corners of the face template from the images in the database (see Fig. 6).

Since the information associated with the appearance of the facial expression is represented by parameters $\mathbf{c}_d$, the expression in the sequence of images $\mathtt{I}_1, \ldots, \mathtt{I}_m$ can be identified as a trajectory, $\mathbf{c}_{d,t}$, $t = 1 \ldots m$, in the

Fig. 4 A picture of a subject (*left* image) compared to the mean face of the model (*right* image)

Fig. 5 Resulting pictures (*right column*) generated by exchanging the mean image in the appearance model with the image shown in the *left column*. In the *middle column*, we show the actual facial expression that we wanted to generate. *Upper* and *lower rows* correspond respectively to individuals 52 and 111 in the Cohn–Kanade database

Fig. 6 Face template used in the construction of the facial expression manifold

deformation subspace. Trajectories associated with the same prototypic facial expression represent roughly similar facial deformations and, consequently, will be located in nearby positions in the deformation subspace. Conversely, the trajectories of different expressions will be located in different positions in the subspace. Figure 7 shows the trajectories of two prototypic facial expressions for three different subjects. We find that the final part of the trajectories of the facial expressions, during the apex, are clearly located in different positions in the deformation subspace. The initial part of all trajectories, associated with the neutral expression, merge in the centre of the plot.

Our model of a prototypic facial expression is the manifold that contains the set of trajectories of that expression in the database. Since all expressions are defined in the

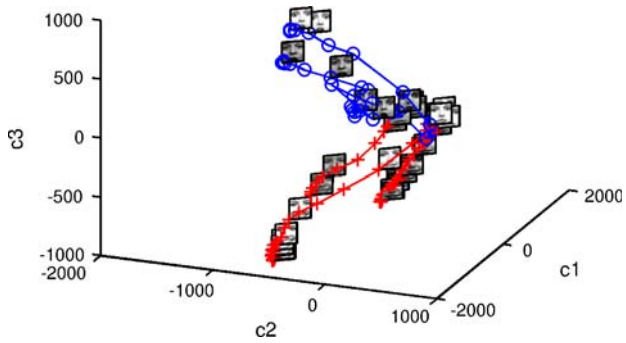


Fig. 7 Trajectories of two prototypic facial expressions (happiness and surprise) for three different subjects in the subspace spanned by the three directions of $\mathtt{B}_d$ with the largest variance. We mark the samples in the happiness and surprise sequences with *crosses* and *circles*, respectively. The marks of each subject are joined by linear segments

common linear space spanned by $\mathtt{B}_d$, our facial expression model is the union of the six manifolds associated with each prototypic facial expression. All six manifolds would merge in the centre of the model, since the initial part of all image sequences corresponds to the neutral expression, and would spread in six different directions depending on the facial expression (see Fig. 8). To diminish the size of the final expression manifold, we only represent in it the last six images of each sequence, because they form the most discriminative part of the sequence.

The dimension of the linear subspace spanning the modes of face deformation $(\dim(\mathtt{B}_d))$ is quite high compared with the amount of data available for training (in the experiments conducted in Sect. 6 this dimension is $\dim(\mathtt{B}_d) = 27$). To avoid the curse of dimensionality and achieve a better generalisation with our facial expression classification algorithm, we use a dimensionality reduction

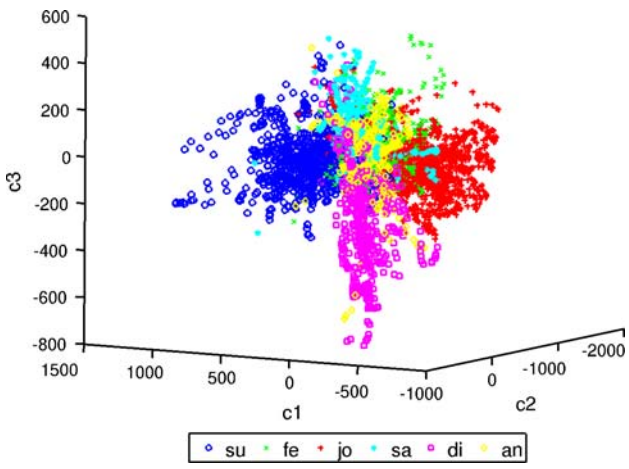


Fig. 8 Facial expression model after reducing the dimensionality to three LDA dimensions. Only the last six images of each sequence are displayed

procedure. Many dimensionality reduction procedures have been introduced in the literature. They can be basically divided into linear and non-linear techniques. Non-linear approaches are the most general but they require a lot of data and often their mappings are defined exclusively on the training data [9, 53, 57]. Linear approaches, on the other hand, are less general, but can be computed with a few data and are defined everywhere in the deformation subspace [8, 30]. In [16] Chang uses the non-linear Lipschitz embedding for dimensionality reduction, whereas Shan uses the linear Locality Preserving Projections (LPP) [30] approach in [56]. For the sake of simplicity, and because it had previously yielded good results [56], we decided to use a linear approach. We chose linear discriminant analysis (LDA) [22] because it performed best in our experiments. In Fig. 8 we show the facial expression model after reducing the dimensionality to three dimensions using LDA.

5 Facial expression recognition

In this section, we introduce a probabilistic facial expression recognition procedure. It combines the prior information stored in the expression manifold with the incoming data obtained from a temporally ordered sequence of images of a face. We recursively estimate the posterior probability of each prototypic facial expression given the incoming image sequence and the set of sequences in the expression manifold. The facial expression in the image sequence is computed as the maximum of the posterior probabilities.

Let $\mathtt{I}_1, \ldots, \mathtt{I}_t$ be a temporally ordered image sequence of a face wearing one or more facial expressions and $\mathbf{x}_1, \ldots, \mathbf{x}_t$ be the temporally ordered set of co-ordinates of the image sequence in the facial expression subspace, which we will denote $\mathcal{X}_{1:t}$. Let $G_t = \{g_1, g_2, \ldots, g_c\}$ be a discrete random variable representing the facial expression at time t and X_t be a continuous random variable associated with the co-ordinates in the facial expression subspace of the image acquired at time t. We will denote by $P(g_i) \equiv P(G_t = g_i)$ the probability that the discrete random variable G_t takes value g_i and by $p(\mathbf{x}) \equiv p(X_t = \mathbf{x})$ the probability density function (p.d.f.) of the continuous variable $\mathbf{x}$ at time t.

The facial expression $g(t)$ at time instant t is obtained as the maximum of the posterior distribution of G_t given the sequence of images up to time t

$$g(t) = \arg\max_i \{P(G_t = g_i|\mathcal{X}_{1:t})\}.$$

Alternatively, the facial expression may also be described as a probabilistic blending of the c primary facial expressions.

We will estimate the posterior distribution using a recursive Bayesian filter. For the first image in the sequence the problem can be immediately solved by

$$P(G_1|\mathbf{x}_1) = \frac{p(\mathbf{x}_1|G_1)P(G_1)}{p(\mathbf{x}_1)} \propto p(\mathbf{x}_1|G_1)P(G_1), \tag{10}$$

where $P(G_1)$ represents our prior knowledge of the probabilities of facial expressions.

Now, if we have a temporal sequence of images $\mathcal{X}_{1:t}$, we can then update G_t as

$$P(G_t|\mathbf{X}_{1:t}) = \frac{p(\mathbf{x}_t|G_t, \mathbf{X}_{1:t-1})p(G_t, \mathbf{X}_{1:t-1})}{p(\mathbf{X}_{1:t})}.$$

If we assume that measurements depend only on the current state, then $p(X_t|G_t, \mathcal{X}_{1:t-1}) = p(X_t|G_t)$ and, hence,

$$P(G_t|\mathcal{X}_{1:t}) \propto p(X_t|G_t)P(G_t|\mathcal{X}_{1:t-1}),$$

where $P(G_t|\mathcal{X}_{1:t-1})$ is the prediction of G_t, given the data up to time instant $t-1$. This probability can be estimated as

$$\begin{aligned} P(G_t|\mathcal{X}_{1:t-1}) &= \sum_{i=1}^{c} P(G_t, G_{t-1} = g_i|\mathcal{X}_{1:t-1}) \\ &= \sum_{i=1}^{c} P(G_t|g_i, \mathcal{X}_{1:t-1})P(g_i|\mathcal{X}_{1:t-1}). \end{aligned}$$

If we assume that our system is Markovian (G_t depends only on G_{t-1}), then

$$P(G_t|\mathcal{X}_{1:t-1}) = \sum_{i=1}^{c} P(G_t|G_{t-1} = g_i)P(G_{t-1} = g_i|\mathcal{X}_{1:t-1}),$$

where $P(G_t|G_{t-1})$ is the expression transition probability.

In contrast to previous approaches (e.g. [16, 56]), which try to estimate the probability of transition between two facial expressions, we believe that all expression transitions are equally probable and introduce the following definition:

$$P(G_t = g_j|G_{t-1} = g_i) = \begin{cases} h & \text{if} \quad j = i \\ \frac{1-h}{c-1} & \text{if} \quad j \neq i, \end{cases} \tag{11}$$

where $0 \leq h \leq 1$ is a smoothing parameter that controls how G_{t-1} influences the predictions about G_t (see Fig. 9). If $h = 1$ no smoothing is performed in the prediction and $P(G_t|G_{t-1}) = P(G_{t-1})$. When $\frac{1}{c} < h < 1$ different degrees of smoothing are performed on $P(G_{t-1})$ to estimate $P(G_t)$. In the extreme case of $h = \frac{1}{c}$, the smoothing is the strongest and $P(G_t|G_{t-1})$ is a uniform distribution ($P(G_t|G_{t-1}) = \frac{1}{c}$). When $0 \leq h < \frac{1}{c}$ smoothing is inverse. In this case expressions that were most probable at $t-1$ are the least probable at t.

In our recognition system, the parameter h acts as a forgetting factor. The closer h is to 1, the less we forget

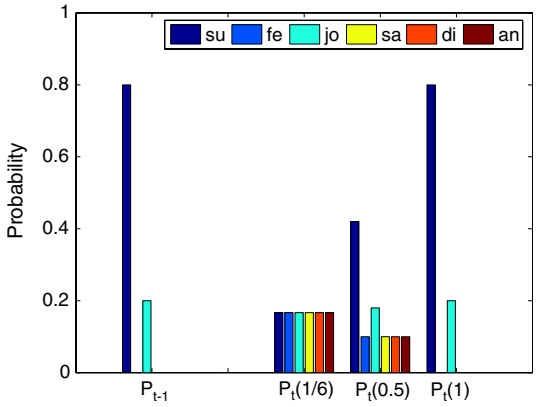


Fig. 9 Effect of parameter h in $p(G_t|\mathcal{X}_{t-1})$. Here P_{t-1} stands for $P(G_{t-1}|\mathcal{X}_{t-1})$, and $P_t(h_i)$ means $P(G_t|\mathcal{X}_{t-1})$ using $h = h_i$

about the information provided by all previous images in the sequence. In extreme cases, when $h = 1$, all images in the sequence are taken into account, and when $h = \frac{1}{c}$, the recognition is performed exclusively on the basis of the last image acquired.

5.1 Estimating $p(X|G)$

$p(\mathbf{x}|g_i)$ represents the p.d.f of an image $\mathtt{I}$ with co-ordinates $\mathtt{x}$ when the subject is wearing facial expression g_i. Our objective here is to estimate this p.d.f. from the data in the facial expression manifold. We will use a k-nearest neighbour approach. Let k be the number of elements in the nearest neighbour set of $\mathbf{x}$, $k_i(\mathbf{x})$ the number of elements in the nearest neighbour set that belong to facial expression g_i ($k = \sum_{i=1}^{c} k_i(\mathbf{x})$) and n_i the number of samples in the manifold of facial expression g_i. Then

$$p(\mathbf{x}|g_i) = \frac{k_i(\mathbf{x})}{n_i \mathcal{V}(k)},$$

where $\mathcal{V}(k)$ is the volume of the neighbourhood enclosing the k nearest neighbours.

The above estimation suffers from the so-called veto effect [2]. If there is a single image in the sequence, $\mathtt{I}_r$, such that $k_i(\mathbf{x}_r) = 0$, then $P(g_i|\mathcal{X}_{1:t}) = 0$, no matter what the values of this probability for all preceding time instants were. This is an undesired event that often occurs when the face is at the apex of an expression. We then introduce a regularised estimation for k_i termed k_i^r such that

$$k_i^r(\mathbf{x}) = \begin{cases} \eta & \text{if} \quad k_i(\mathbf{x}) = 0, \\ k_i(\mathbf{x}) & \text{otherwise}, \end{cases}$$

where the parameter $0 \leq \eta \leq 1$ models the amount of regularisation introduced for a facial expression with no neighbour. We also normalise $k_i^r(\mathbf{x})$ such that $\sum_{i=1}^{c} k_i^r(\mathbf{x}) = k$. So, we estimate $p(X|G)$ as

$$p(\mathbf{x}|g_i) = \frac{k_i^r(\mathbf{x})}{n_i \mathcal{V}(k)} \propto \frac{k_i^r(\mathbf{x})}{n_i}.$$

6 Experiments

In this section, we evaluate the performance of the facial expression recognition system described in this paper. We have conducted two groups of experiments. The objective of the first set of experiments is to qualitatively validate the performance of the system by comparing the results obtained by the facial expression recognition procedure with our subjective classification. In the second group of experiments, we quantitatively test the performance of the system by classifying the 333 sequences from the Cohn–Kanade database used to build the expression manifold.

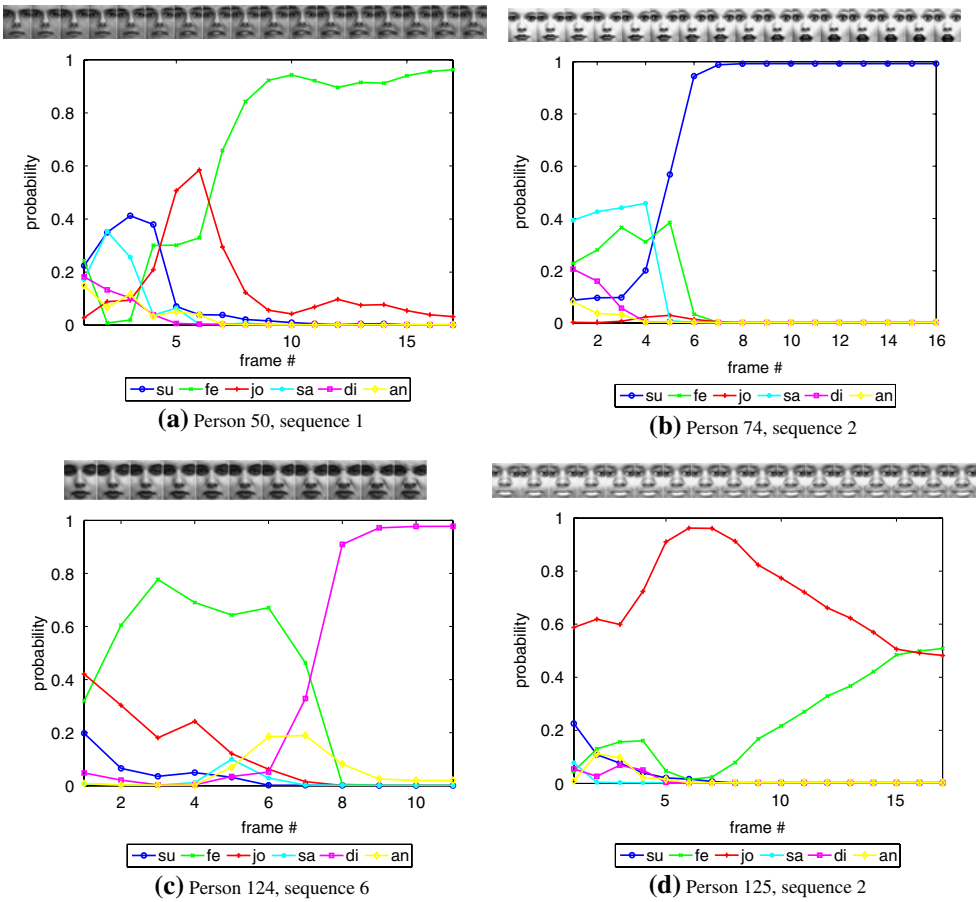
The linear subspace of the tracker used in this section was obtained with the procedure and the training data described in Sect. 3.3. The dimension of the deformation subspace which results from the training process is $\dim(\mathtt{B}_d) = 27$.

In all experiments, we followed a cross-validation scheme: each sequence was classified eliminating all other sequences for the same subject from the facial expression model. We have also assumed that all facial expressions are equally probable, e.g. $P(G_1)$ in (10) is the same for all facial expressions.

6.1 Qualitative experiments

With these experiments, we analyse various image sequences and compare the evolution of the recognition process in the system with our subjective impression.

For the first experiment, we have selected four test sequences from the Cohn–Kanade database. The dimension of the deformation subspace of the tracker was reduced to 5 using LDA. The number of nearest neighbours used to estimate $p(X|G)$ is 31 and parameter h has a value of 0.3. Figure 10 shows the results of the recognition process for the test sequences. In the first sequence, shown in Fig. 10a, the true facial expression is fear and the system correctly recognises it. From frame seven onwards the motion of the mouth and eyebrows is the movement associated with fear. Before this point, motion is only associated with the mouth, and other facial expressions (surprise and joy) are recognised. A similar thing applies to the surprise expression in Fig. 10b, where the eyebrows start to rise in frame five. The most discriminative feature of the expression of disgust is frowning in the face region between the eyebrows



(a) Person 50, sequence 1

(b) Person 74, sequence 2

(c) Person 124, sequence 6

(d) Person 125, sequence 2

Fig. 10 Classification experiment using four sequences from the Cohn–Kanade database

and the nose. This clearly happens from frame seven onwards in Fig. 10c. Before this frame, the expression may be confused with sadness because of the shape of the mouth and eyebrows. Finally, the expression in Fig. 10d was labelled as joy and classified by our system as fear. In this case the expression the subject wears is unclear and it is difficult even for us to assign it to an expression class. Nevertheless, since our classifier is probabilistic, we can see that the probabilities of joy and fear are very similar.

For the second qualitative experiment, we acquired a sequence in which a talking face wears three expressions (joy, surprise and anger) in a realistic situation with varying illumination and face motion (see Fig. 11). For this test the model included the neutral facial expression. The results of the recognition process are shown in Fig. 12. From frames 0 to 300 the actor is moving, talking and wearing one facial expression (joy in frame 39). In this part of the sequence, the actor also wears several expressions that do not directly correspond to the any of the facial expressions in the model. For example, frame 231 is almost a joy expression, but no teeth were displayed, and the eyebrows are raised in frame 296. From frames 290 to 805, the motion of a tungsten light produces sharp changes in the illumination of the face. As we will see, system performance is not severely affected by these changes. This is thanks to a correct tracker performance, whose illumination subspace absorbs these variations in most of this part of the sequence. Between frames 300 and 500, there are three surprise and one joy expressions, worn in varying positions and with small out-of-plane head motions. They are correctly recognised. From frames 530 to 650 we have an anger expression, which is correctly recognised in spite of strong translational and small out-of-plane head motion. Finally, the surprise and joy expressions in frames 859 and 930 are also correctly recognised.

In some situations, the system does not give a correct classification. This is because of expressions not represented in our model, like the tongue sticking out in frame 482. Other failures are caused by tracking inaccuracies, such as the surprise expressions in frames 689 and 820.

6.2 Quantitative experiments

Here, we quantitatively evaluate the performance of our facial expression recognition algorithm for different configuration parameters and dimensionality reduction procedures. The performance of the best configuration will

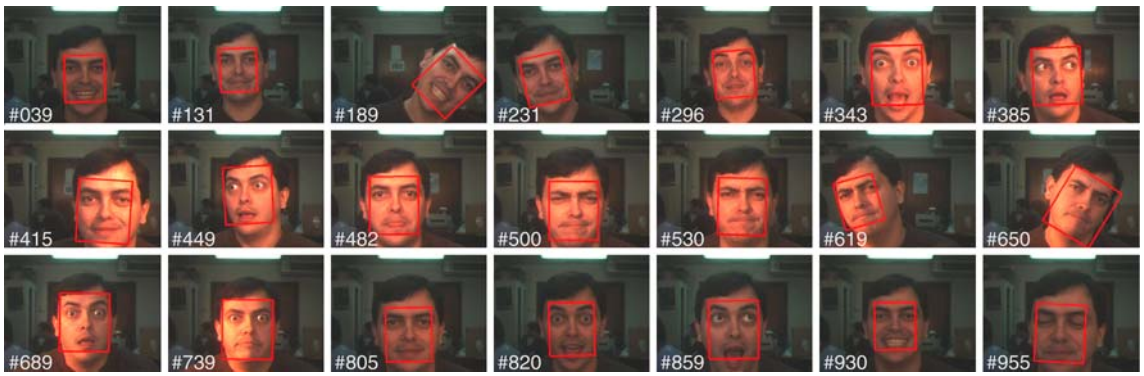


Fig. 11 Tracking results for a realistic sequence

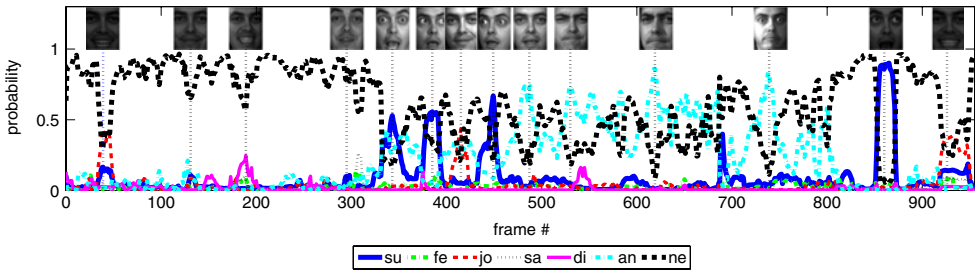


Fig. 12 Facial expression recognition in a realistic image sequence

then be compared with other recognition systems. For our tests we will use once again the 333 manually labelled image sequences from Cohn–Kanade database used to build the facial expression model.

In the first experiment, we test the performance of our classification algorithm with two linear dimensionality reduction procedures: LDA [22] and the supervised version of LPP [30] introduced in [55]. In this case, our baseline classifier (the one with no dimensionality reduction) uses the raw facial deformation parameters coming from the appearance-based tracker. This is equivalent to a principal component analysis (PCA) [22] projection of the incoming image pixel intensities. Classification success for a sequence is declared whenever the posterior probability of the true facial expression is the largest at the last frame of the sequence. Tables 1, 2 and 3 show the confusion matrices resulting from the classification of the 333 test sequences using the baseline, LPP and LDA classifiers, respectively.

From these results, we can conclude that, as expected, supervised dimensionality reduction approaches (LDA and supervised LPP) achieve better recognition rates than PCA. We selected LDA as the dimensionality reduction procedure for our system, since it achieved a marginal improvement over the supervised LPP. Another conclusion is that surprise and joy are the easiest facial expression to recognise, since they involve strong appearance variations: open mouth and raised eyebrows for surprise, and open mouth and displayed teeth for joy. On the other hand, fear, sadness and anger are the most difficult expressions to recognise because they involve more subtle changes in appearance.

Table 1 Confusion matrix (expressed in percentage) for the baseline classification experiment

	su	fe	jo	sa	di	an	Total
su	91.43	2.38	0	0	2.44	0	
fe	4.29	59.52	2.44	12	7.32	10.81	
jo	2.85	21.43	97.56	2	2.44	5.4	
sa	1.43	9.52	0	80	4.88	13.51	
di	0	0	0	0	78.05	10.81	
an	0	7.14	0	6	4.88	59.46	
Total							81.67

Figure 13 plots the classification rates achieved using five LDA dimensions while varying both parameter h and the number of nearest neighbours, k, used to estimate $p(X|G)$. The performance of the system grows very fast for values of k between 0 and 10. Between 10 and 40, it grows at slower pace. Beyond that value it does not grow at all. The best performance is achieved for a value of $h = 0.3$ ($k = 31$), although differences are almost negligible for values of h between 0.16 and 0.8. Values of h close to 1 achieve the worst performances. This behaviour is due to the special structure of the sequences in the Cohn–Kanade database, all of which start with a neutral expression and finish at the expression apex. In consequence most of the

Table 2 Confusion matrix (expressed in percentage) for the LPP classification experiment

	su	fe	jo	sa	di	an	Total
su	100	2.38	0	0	0	2.7	
fe	0	73.81	1.22	6	0	2.7	
jo	0	9.52	98.78	4	4.88	2.7	
sa	0	9.52	0	84	7.32	8.11	
di	0	0	0	6	80.49	10.81	
an	0	4.76	0	0	7.32	72.97	
Total							88.20

Table 3 Confusion matrix for the LDA classification experiment

	su	fe	jo	sa	di	an	Total
su	100	0	0	0	0	2.7	
fe	0	73.9	1.2	4	0	0	
jo	0	9.5	98.8	4	0	0	
sa	0	9.5	0	82	4.8	5.4	
di	0	0	0	6	87.9	13.5	
an	0	7.1	0	4	7.3	78.4	
Total							89.13

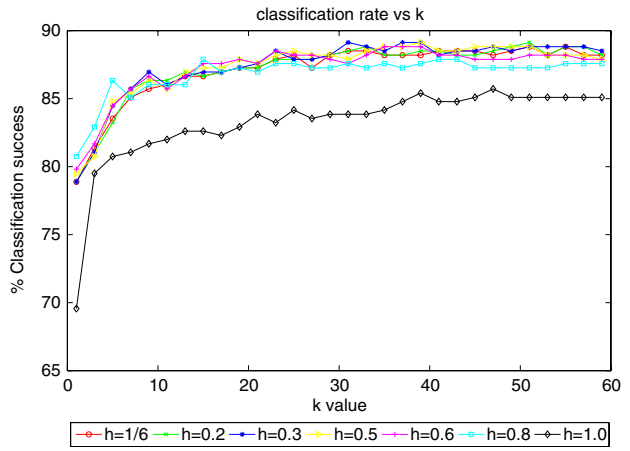


Fig. 13 Classification rate for different values of nearest neighbours using five LDA dimensions

discriminative information is stored in the last frames of the sequences, near the expression apex. As the value of h grows closer to 1, more frames of the initial part of the sequence are considered in the computation of the posterior probabilities. Consequently, performance decreases, since the initial frames have similar appearances for all facial expressions.

Table 4 lists the recognition results of our system together with other results previously described in the literature. Unfortunately, these results cannot be directly compared because they were obtained with different data. Michel and El Kailouby [40] use a set of 72 test examples from a subject familiar with their system. Although the other five systems are based on the Cohn–Kanade database, they use different sequences. Of the 485 image sequences of 97 individuals in the database, we use 333, Zhao [70] and Shan [56] use, respectively, 374 and 316, and finally Cohen [17] uses sequences related to 53 individuals. Moreover, even if all systems had used the same sequences, the labelling could be different, since the translation from FACS AUs to the primary expression may not be standard. A common set of sequences with associated labels is necessary to make fair comparisons. The sequences that we used in this paper and their labels are publicly available at http://www.dia.fi.upm.es/pcr/face_expressions.htm.

From Table 4 we can conclude that our system's performance is similar to some of the best performing systems (Shan06 and Yeasin04), although Zhao's results are clearly ahead. Nevertheless, our system is able to work in a realistic set up with sharp illumination variations, small rotations of the face out of camera plane and large in-plane rotations and translations.

7 Conclusions

In this paper, we have introduced a system that recognises facial expressions in video sequences. It uses the deformation parameters provided by a dense and efficient appearance-based face tracker. The tracker is able to run at standard video frame rates and is robust to illumination variations.

We have also introduced a model-based facial expression recognition system. A facial expression is represented by a set of samples that model a low dimensional manifold in the space of deformations generated by the tracker parameters. In our approach, an image sequence becomes a path in the space of deformations. We use a nearest-neighbour technique to estimate the probability of occurrence of an image from the facial expression sequence. Finally, with a recursive Bayesian procedure, we sequentially combine these probabilities to estimate a posterior for each facial expression. A target sequence may be assigned to the facial expression with maximum posterior probability, or it may also be described as a probabilistic blending of primary facial expressions. Our solution differs from previous related approaches [16, 56] in various ways: (a) our manifolds are user independent, while those introduced in [16] depend on the identity of the user; (b) we use the parameters of an illumination-independent linear model as discriminative information for expression classification, whereas active wavelet networks [31] and local binary

Table 4 Comparing the performance of our system

Reference Approach	Zhao07 [70] SVM	Yeasin04 [66] HMM	Cohen03 [17] TAN	Shan06 [56] Bayesian	Michel03 [40] SVM	Our method Bayesian
Surprise	98.6	100	93.3	98.8	100	100
Fear	94.6	76.4	63.8	66.7	83.3	73.9
Joy	96.0	96.6	86.2	100	75.0	98.8
Sadness	95.8	96.2	61.2	81.7	83.3	82
Disgust	94.7	62.5	62.2	97.5	100	87.9
Anger	96.8	100	66.4	84.2	83.3	78.4
Neutral	–	–	78.5	–	–	–
Total	96.2	90.9	73.2	91.8	87.5	89.13

pattern features [43] are used, respectively, in [16] and [56]; (c) our algorithm for estimating the posterior probabilities is different from those in [16] and in [56] regarding both the estimation of $p(X|G)$ and the assumption that all transitions between facial expressions are equally probable. Here we introduce a function $p(G_t|G_{t-1})$. This function depends on a parameter h that models the size of the temporally ordered set of images used to predict $p(G_t|\mathcal{X}_{t-1})$.

Thanks to the robustness of the tracker, our system is able to work in a realistic set up with sharp illumination variations, small rotations of the face out of camera plane and large in-plane rotations and translations. In the future we may achieve small improvements in the performance of our system using a more involved dimensionality reduction technique and introducing a post-classification procedure to refine the system's decision in difficult sequences. Larger improvements of performance will come mainly from the integration of multiple modalities, such as voice analysis and context.

Both the face tracker and the model of facial expressions can be easily reconfigured. Although the tracker has a user-independent working mode, it can also be configured to work in a user-dependent mode which provides better accuracy. The training for either mode is fully automatic. In our facial expression model, we have introduced Ekman's six prototypic facial expressions, but any other set of expressions could be used by just tracking a set of sample sequences and introducing those samples in a new expression manifold.

In spite of the existence of the Cohn–Kanade database, the scientific community is unable to make fair comparisons of facial expression analysis systems because there is no agreed upon set of sequences and labels. We contribute to the solution of this problem by publishing the sequences and labellings used in our experiments.

Acknowledgments The authors gratefully acknowledge funding from the Spanish *Ministerio de Educación y Ciencia* under contract TRA2005-08529-C02-02. They also thank the anonymous reviewers for their comments and Jeffrey Cohn and Takeo Kanade for providing the Cohn--Kanade image database.

References

1. Flanagan JL, Huang TS (2003) Special issue on human–computer multimodal interface. Proc IEEE, 91(9):1267–1468
2. Alkoot FM, Kittler J (2002) Moderating k-nn classifiers. Pattern Anal Appl 5:326–332
3. Baker S, Matthews I , Schneider J (2004) Automatic construction of active appearance models as an image coding problem. IEEE Trans Pattern Anal Mach Intell 26(10):1380–1384
4. Barlett MS, Littlewort G, Braathen B, Sejnowki T, Movellán J (2003) A prototype for automatic recognition of spontaneous facial actions. In: Becker S, Obermayer K (eds) Advances in neural information processing systems, vol 15. MIT Press, Cambridge, pp 1271–1278
5. Barlett MS, Littlewort G, Frank M, Lainscsek C, Fasel IR, Movellán J (2005) Recognizing facial expression: machine learning and application to spontaneous behaviour. In: Proceedings of CVPR, vol 2, pp 568–573
6. Bascle B, Blake A (1998) Separability of pose and expression in facial tracing and animation. In: Proceedings of international conference on computer vision. IEEE, Washington, pp 323–328
7. Basili JN (1979) Emotion recognition: the role of facial movement and the relative importance of upper and lower area of the face. J Pers Soc Psychol 37:2049–2059
8. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
9. Belkin M, Niyogi P (2001) Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in neural information processing systems, pp 585–591
10. Black MJ, Jepson AD (1998) Eigentracking: robust matching and tracking of articulated objects using a view-based representation. Int J Comput Vis 26(1):63–84
11. Black MJ, Yacoob Y (1997) Recognizing facial expressions in image sequences using local parameterized models of image motion. Int J Comput Vis 25(1):23–48
12. Blanz V, Vetter T (1999) A morphable model for the synthesis of 3d faces. In: Proceedings of SIGGRAPH. ACM Press, New York, pp 187–194
13. Buenaposada JM (2001) Buenaposada and Luis Baumela: variations of grey world for face tracking. Image Process Commun 7(3, 4):51–61
14. Buenaposada JM (2002) Buenaposada and Luis Baumela: real-time tracking and estimation of plane pose. In: Proceedings of

international conference on pattern recognition, vol II, QC, Canada, August 2002. IEEE, Washington, pp 697–700
15. Buenaposada JM, Muñoz E, Baumela L (2006) Efficiently estimating facial expression and illumination in appearance-based tracking. In: Proceedings British machine vision conference, vol I, pp 57–66
16. Chang Y, Hu C, Turk M (2004) Probabilistic expression analysis on manifolds. In: Proceedings of CVPR, vol 2, pp 520–527
17. Cohen I, Sebe N, Garg A, Chen LS, Huang TS (2003) Facial expression recognition from video sequences: temporal and static modeling. Comput Vis Image Underst 91:160–187
18. Cootes T, Edwards GJ, Taylor C (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(6):681–685
19. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human–computer interaction. Signal Process Mag, 18(1):32–80
20. DeCarlo D, Metaxas D (2000) Optical flow constraints on deformable models with applications to face tracking. Int J Comput Vis 38(2):99–127
21. DeCarlo D, Metaxas D(2000) Optical flow constraints on deformable models with applications to face tracking. Int J Comput Vis 38(2):99–127
22. Duda RO, Hart PE, Stork DG (2000) Pattern classification. Wiley, New York
23. Ekman P (1993) Facial expression and emotion. Am Psychol 44:384–392
24. Ekman P (1994) Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. Psychol Bull 115(2):268–287
25. Essa I, Pentland A (1997) Coding, analysis, interpretation, recognition of facial expressions. IEEE Trans Pattern Anal Mach Intell 19(7):757–763
26. Fasel B, Luettin J (2003) Automatic facial expression analysis: a survey. Pattern Recognit 36:259–275
27. Gao Y, Leung MKH, Hui SC, Tanada MW (2003) Facial expression recognition from line-based caricatures. Trans SMC A 33(3):407–412
28. Gee A, Cipolla R (1996) Fast visual tracking by temporal consensus. Image Vis Comput 14(2):105–114
29. Hager G, Belhumeur P (1998) Efficient region tracking with parametric models of geometry and illumination. IEEE Trans Pattern Anal Mach Intell 20(10):1025–1039
30. He X, Niyogi P (2003) Locality preserving projections. In: Thrun S, Saul L, Schölkopf B (eds) Advances in neural information processing systems, vol 16. MIT Press, Cambridge
31. Hu C, Ferris R, Turk M (2003) Active wavelet networks for face alignment. In: Proceedings of British machine vision conference
32. Jongwoo L, Ross D, Ruei-Sung L, Ming-Hsuan Y (2004) Incremental learning for visual tracking. In: Advances in neural information processing systems, vol 17. MIT Press, Cambridge, 793–800
33. Kanade T, Cohn J, Tian Y-L (2000) Comprehensive database for facial expression analysis. In: Proceedings of international conference on automatic face and gesture recognition, pp 46–53
34. Lanitis A, Taylor CJ, Cootes TF (1997) Automatic interpretation and coding of face images using flexible models. IEEE Trans Pattern Anal Mach Intell 19(7):743–756
35. Lien JJ, Kanade T, Cohn JF, Li C (1997) Detection, tracking and classification of action units in facial expression. J Rob Auton Syst 31:131–146
36. Lyons MJ, Budynek L, Akamatsu S (1999) Automatic classification of single facial images. IEEE Trans Pattern Anal Mach Intell 21(12):1357–1362
37. Mase K (1991) Recognition of facial expression from optical flow. IEICE Trans E 74(10):3474–3483
38. Matthews I, Baker S (2004) Active appearance models revisited. Int J Comput Vis 60(2):135–164
39. McTear MF (2002) Spoken dialogue technology: enabling the conversational user interface. ACM Comput Surv 34(1):90–169
40. Michel P, El Kaliouby R (2003) Real time facial expression recognition in video using support vector machines. In: Proceedings of international conference on multimodal interfaces. ACM, New York, pp 258–264
41. Muñoz E, Buenaposada JM, Baumela L (2005) Efficient model-based 3d tracking of deformable objects. In: Proceedings of international conference on computer vision, Beijing, China, vol I, pp 877–882,
42. Ohya J, Kitamura Y, Takemura H, Ishi H, Kishino F, Terashima N (1996) Virtual space teleconferencing: real-time reproduction of 3d human images. J Vis Commun Image Represent 6(1):1–25
43. Ojala T, Pietikainen M, Menpp T (2002) Multirresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
44. Oliver N, Pentland A, Bérard F (2000) Lafter: a real-time face and lips tracker with facial expression recognition. Pattern Recognit 33:1369–1382
45. Panti M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. IEEE Trans Pattern Anal Mach Intell 22(12):1424–1445
46. Pantic M, Rothkrantz LJM (2000) Expert system for automatic analysis of facial expressions. Image Vis Comput 18(11):881–905
47. Picard RW (1997) Affective computing. MIT Press, Cambridge
48. Raducanu B, Graña M, Albizuri FX, d'Anjou A (2004) A probabilistic hit-and-miss transform for face localization. Pattern Anal Appl 7:117–127
49. Rani P, Liu C, Sarkar N (2006) An empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Anal Appl 9:58–69
50. Romdhani S, Vetter T (2003) Efficient robust and accurate fitting of a 3rd morphable model. In: Proceedings of international conference on computer vision, vol 1, pp 59–66
51. Rose N (2006) Facial expression classification using gabor and log-gabor filters. In: Proceedings of international conference on automatic face and gesture recognition
52. Rosenblum M, Yacoob Y, Davis LS (1996) Human expression recognition from motion using radial basis function network architecture. IEEE Trans Neural Netw 7(5):1121–1138
53. Roweis S, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
54. Rowley H, Baluja S, Kanade T (1998) Neural network-based face detection. IEEE Trans Pattern Anal Mach Intell 20(1):23–28
55. Shan C, Gong S, McOwan PW (2005) Appearance manifold of facial expression. In: IEEE international workshop on human–computer interaction
56. Shan C, Gong S, McOwan PW (2006) Dynamic facial expression recognition using a bayesian temporal manifold model. In: Proceedings of British machine vision conference vol 1, pp, 297–306
57. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
58. Terzopoulos D, Waters K (1993) Analysis and synthesis of facial image sequences using physical and anatomical models. IEEE Trans Pattern Anal Mach Intell 15(6):569–579
59. Tian Y, Kanade T, Cohn J (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23(2):97–115
60. Tong Y, Liao W, Ji Q (2006). Inferring facial action units with causal relations. In: Proceedings of CVPR , vol 2, pp 1623–1630
61. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86

62. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
63. Wang J, Yin L, Wei X, Sun Y (2006) 3d facial expression recognition based on primitive surface feature distribution. In: Proceedings of CVPR, vol 2, pp 1399–1406
64. Xu Y, Roy-Chowdhury AK (2007) Integrating motion, illumination and structure in video sequences with applications in illumination-invariant tracking. IEEE Trans Pattern Anal Mach Intell 29(5):793–806
65. Yacoob Y, Davis LS (1996) Recognizing human facial expressions from long image sequences using optical flow. IEEE Trans Pattern Anal Mach Intell 18(6):636–642
66. Yeasin M, Bullot B, Sharma R (2004) From facial expression to levels of interest: a spatio-temporal approach. In: Proceedings of CVPR, vol 2, 922–927
67. Zhang Y, Ji Q (2003) Facial expression understanding in image sequences using dynamic and active information fusion. In: Proceedings of international conference on computer vision, Nice, France
68. Zhang Y, Ji Q (2005) Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans Pattern Anal Mach Intell 27(5):1–16
69. Zhang Z, Lyons M, Schuster M, Akamatsu S (1998) Comparison between geomtry-based and gabor wavelets-based facial expression recognition using multi-layer perceptron. In: Proceedings of international conference on automatic face and gesture recognition, Nara, Japan, pp 454–459
70. Zhao G, Pietikäinen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928