A Dynamic Texture based Approach to Recognition of Facial Actions and their Temporal Models

Sander Koelstra, *Student Member IEEE*, Maja Pantic, *Senior Member IEEE*, Ioannis Patras, *Member IEEE*

Abstract—In this work we propose a dynamic-texture-based approach to the recognition of facial Action Units (AUs, atomic facial gestures) and their temporal models (i.e., sequences of temporal segments: neutral, onset, apex, and offset) in near-frontal-view face videos. Two approaches to modelling the dynamics and the appearance in the face region of an input video are compared: an extended version of Motion History Images and a novel method based on Non-rigid Registration using Free-Form Deformations (FFDs). The extracted motion representation is used to derive motion orientation histogram descriptors in both the spatial and temporal domain. Per AU, a combination of discriminative, frame-based GentleBoost ensemble learners and dynamic, generative Hidden Markov Models detects the presence of the AU in question and its temporal segments in an input image sequence. When tested for recognition of all 27 lower and upper face AUs, occurring alone or in combination in 264 sequences from the MMI facial expression database, the proposed method achieved an average event recognition accuracy of 89.2% for the MHI method and of 94.3% for the FFD method. The generalization performance of the FFD method has been tested using the Cohn-Kanade database. Finally, we also explored the performance on spontaneous expressions in the Sensitive Artificial Listener dataset.

Index Terms—facial image analysis, facial expression, dynamic texture, motion

1 Introduction

A widely accepted prediction is that computing will move to the background, weaving itself into the fabric of our everyday living and projecting the human user into the foreground [1]. To realise this goal, next-generation computing (a.k.a. pervasive computing, ambient intelligence, human computing) will need to develop human-centred user interfaces that respond readily to naturally occurring, multi-modal, human communication [24]. These interfaces will need the capacity to perceive and understand human users' intentions and emotions as communicated by social and affective signals. Motivated by this vision of the future, automated analysis of non-verbal behaviour, and especially of facial behaviour, has attracted increasing attention in computer vision, pattern recognition, and human-computer interaction. Facial expression is one of the most cogent, naturally pre-eminent means for human beings to communicate emotions, to clarify and stress what is said, to signal comprehension, disagreement, and intentions, in brief, to regulate interactions with the environment and other persons in the vicinity [11]. Automatic analysis of facial expressions forms, therefore, the essence of numerous next-generation-computing tools including affective computing technologies (i.e. proactive and affective user interfaces), learner-adaptive tutoring systems, patient-profiled personal wellness technologies, etc. [21]. In general, since facial expressions can predict the onset and remission of depression and schizophrenia, certain brain lesions, transient myocardial ischemia, different types of pain (acute vs. chronic), and can help identify alcohol intoxication and deception, the potential benefits from efforts to automate the analysis of facial expressions are varied and numerous and span fields as diverse as cognitive sciences, medicine, education, and security [21].

Two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) detection and facial muscle action (action unit) detection [25, 21, 41]. The most commonly used facial expression descriptors in facial affect detection approaches are the six basic emotions (fear, sadness, happiness, anger, disgust, surprise), proposed by Ekman and discrete emotion theorists, who suggest that these emotions are universally displayed and recognized from facial expressions. The most commonly used facial muscle action descriptors are the Action Units (AUs) defined in the Facial Action Coding System (FACS; [10]).

This categorization in terms of six basic emotions used in facial affect detection approaches, though quite intuitive, has some important downsides. The basic emotion categories form only a subset of the total range of possible facial displays and categorization of facial expressions can therefore be forced and unnatural. Boredom and interest, for instance, do not seem

- *Sander Koelstra (sander.koelstra@elec.qmul.ac.uk) and Ioannis Patras (i.patras@elec.qmul.ac.uk) are with Queen Mary University of London, E14NS, UK.*
- *Maja Pantic (m.pantic@imperial.ac.uk) is with Imperial College London, SW72AZ, UK, and with University of Twente, 7500 AE Enschede, NL.*
- *The authors would like to thank Jeffrey Cohn of the University of Pittsburgh for providing the Cohn-Kanade database. The research of Sander Koelstra has received funding from the Seventh Framework Programme under grant agreement no. FP7-216444 (PetaMedia). This work has been funded first in part by the ECs 7th Framework Programme [FP7 / 2007-2013] under grant agreement no 211486 (SEMAINE). Current research of Maja Pantic is funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The research of Ioannis Patras has been partially supported by EPSRC Grant No EP/G033935/1.*

Fig. 1: Apex phases of 8 AUs of the FACS system.

to fit well in any of the basic emotion categories. Moreover, in everyday life, these prototypic expressions occur relatively rarely; usually, emotions are displayed by subtle changes in discrete facial features, such as raising of the eyebrows in surprise. To detect such subtlety of human emotions and, in general, to convey the information on facial expressions to aforementioned applications, automatic recognition of atomic facial signals, such as the AUs of the FACS system, is needed.

FACS was proposed by Ekman and Friesen in 1978 and revised in 2002 [10]. FACS classifies atomic facial signals into Action Units (AUs) according to the facial muscles that cause them. It defines 9 upper face AUs and 18 lower face AUs, which are considered to be the smallest visually discernible facial movements. It also defines 20 Action Descriptors for eye and head position. FACS provides the rules both for AU intensity scoring and for recognition of temporal segments (onset, apex and offset) of AUs in a face video.

Most of the research on automatic AU recognition has been based on analysis of static images (e.g. [26]) or individual frames of an image sequence (e.g. [3, 4, 18, 17]). Some research efforts toward using dynamic textures (DT) for facial expression recognition (e.g. [36, 43]) and toward explicit coding of AU dynamics (e.g. with respect to AUs temporal segments, like in [23, 35], or with respect to temporal correlation of different AUs like in [33]) have been proposed as well. However, most of these previously proposed systems recognise either the six basic emotions (e.g. [43]) or only subsets of the 27 defined AUs. Except for geometric-feature-based methods proposed in [22, 23, 35], none of the existing systems attains automatic recognition of AUs temporal segments. Also, except for the method based on Motion History Images proposed in [36], none of the past works attempted automatic AU recognition using of a DT-based approach.

In this work we present a novel DT-based approach to automatic facial expression analysis in terms of all 27 AUs and their temporal segments. The novelties in this work are:

- We propose a new set of adaptive and dynamic texture features for representing facial changes that are based on Free-form Deformations (FFD).
- We introduce a novel non-uniform decomposition of the facial area to facial regions within which features are extracted. This is based on a quadtree decomposition of motion images, and results in more features being allocated to areas that are important for recognition of an AU and less features being allocated to other areas.
- We combine a discriminative, frame-based GentleBoost classifier with a dynamic, generative HMM model for (temporal) AU classification in an input face video.
- This is the second DT-based method for AU recognition proposed. We compare our method to the earlier method [36], and show a clear improvement in performance.

An early version of this work appeared in [16]. The outline of the paper is as follows. Section 2 provides an overview of the related research. Section 3 presents the two utilized approaches to modelling dynamics and the appearance in the face region of an input video (MHI and FFD) and explains the methodology used to detect AUs and their temporal segments. Section 4 describes the utilized datasets, the evaluation study and discusses the results. Section 5 concludes the paper.

2 State of the Art

2.1 Facial Features

Existing approaches to facial expression analysis can be divided into geometric and appearance-based approaches. Dynamic texture recognition can be seen as a generalization of appearance-based approaches. Geometric features include shapes and positions of face components, as well as the location of facial feature points (such as the corners of the mouth). Often, the position and shape of these components and/or fiducial points are detected in the first frame and then tracked throughout the sequence. On the other hand, appearance-based methods rely on skin motion and texture changes (deformations of the skin) such as wrinkles, bulges and furrows. Both approaches have advantages and disadvantages. Geometric features only consider the motion of a number of points, so one ignores much information present in the skin texture changes. On the other hand, appearance-based methods may be more susceptible to changes in illumination and differences between individuals. See [25, 40] for an extensive overview of facial expression recognition methods.

2.1.1 Geometric-feature-based approaches

Approaches that use only geometric features mostly rely on detecting sets of fiducial facial points (e.g. [26, 23, 35]), a connected face mesh or active shape model (e.g. [13, 7, 5, 17]), or face component shape parametrization (e.g. [31]). Next, the points or shapes are tracked throughout the video and the utilized features are their relative and absolute position, mutual spatial position, speed, acceleration, etc. A geometric approach that attempts to automatically detect temporal segments of AUs is the work of Pantic and colleagues [22, 23, 35]. They locate and track a number of facial fiducial points and extract a set of spatio-temporal features from the trajectories. In [22] and [23], they use a rule-based approach to detect AUs and their temporal segments, while in [35] they use a combination of SVMs and HMMs to do so. Using only the movement of a number of feature points makes it difficult to detect certain AUs, such as AU 11 (nasolabial furrow deepener), 14 (mouth corner dimpler), 17 (chin raiser), 28 (inward sucking of the lips) (see also Fig. 1), the activation of which is not apparent from movements of facial points but rather from changes in skin texture. Yet, these AUs are typical for facial expressions of emotions such as sadness (see EMFACS [10]), and for expressions of more complex mental states including puzzlement and disagreement [11], which are of immense

importance if the goal is to realize human-centred, adaptive interfaces. On the contrary, our appearance-based approach is capable of detecting the furrows and wrinkles associated with these AUs and is therefore better equipped to recognize them.

2.1.2 Appearance-based approaches

Systems using only appearance-based features have been proposed in e.g. [18, 3, 4, 14, 2, 20, 36]. Several researchers have used Gabor wavelet coefficients as features (e.g. [14, 42, 38]). Bartlett et al. [3, 18, 4] have tried different methods such as optical flow, explicit feature measurement (i.e. length of wrinkles, degree of eye opening), ICA and the use of Gabor wavelets. They report that Gabor wavelets render the best results [18]. Other techniques used include optical flow [2] and Active Appearance Models [20]. Tian et al. [31, 32] use a combination of geometric and appearance-based features (Gabor wavelets). They claim that the former features outperform the latter ones, yet using both yields the best result.

2.1.3 Dynamic-Texture based approaches

An emerging new method of appearance-based activity recognition is known as Dynamic Texture recognition. A Dynamic Texture (DT) can be defined as a "spatially repetitive, time-varying visual pattern that forms an image sequence with certain temporal stationarity" [6]. Typical examples of DTs are smoke, fire, sea waves and talking faces. Many existing approaches to recognition of DTs are based on optical flow [28, 19]. A different approach is used in [30]. Instead of using optical flow, they use system identification techniques to learn generative models. Recently, Chetverikov and Péteri [6] published an extensive overview of DT approaches.

The techniques applied to the DT recognition problem can also be used to tackle the problem of facial expression recognition. Valstar et al. [36] encoded face motion into Motion History Images. This representation shows a sequence of motion energy images superimposed in a single image, detailing recent motion in the face. An extended version of MHI-based facial expression recognition is proposed in this work as well. In this work, videos are temporally segmented by manually selecting the start and endpoints of an AU activation and a single MHI is created from 6 frames distributed equidistantly between these points. In our implementation, an MHI is created for a temporal window around each frame without any manual input. Also, while their method uses a multi-class classifier, we train separate binary classifiers for each AU and therefore we can detect any combination of AUs.

Zhao and Pietikäinen [43, 44] use volume local binary patterns (LBP), a temporal extension of local binary patterns often used in 2D texture analysis. The face is divided into overlapping blocks and the extracted LBP features in each block are concatenated into a single feature vector. SVMs are used for classification. The approach shows promising results, although only the six prototypic emotions are recognized and no temporal segmentation is performed. They normalize the face using the eye position in the first frame, but they ignore any rigid head movement that may occur during the sequence. In addition, instead of our learned class(AU)-specific quadtree placement method for feature extraction regions, they use fixed overlapping blocks distributed evenly over the face. To the best of our knowledge, our method is the only other DT-based method for facial expression analysis proposed so far.

3 METHODOLOGY

Fig. 2 gives an overview of our system. In the preprocessing phase, the face is located in the first frame of an input video and head motion is suppressed by an affine rigid face registration. Next, non-rigid motion is estimated between consecutive frames by the use of either Non-rigid Registration using Free-form Deformations (FFDs) or Motion History Images (MHIs). For each AU, a quadtree decomposition is defined to identify face regions related to that AU. In these regions, orientation histogram feature descriptors are extracted. Finally, a combined GentleBoost classifier and a Hidden Markov Model (HMM) are used to classify the sequence in terms of AUs and their temporal segments. In the remainder of this section the details of each processing phase are described.

3.1 Rigid face registration

In order to locate the face in the first frame of the sequence, we assume the face is expressionless and in a near-frontal position in that frame and use the fully automatic face and facial point detection algorithm proposed in [37]. This algorithm uses an adapted version of the Viola-Jones face detector to locate the face. 20 facial characteristic points and a facial bounding box are detected by using Gabor-feature-based boosted classifiers.

To suppress inter-sequence variations (i.e. facial shape differences) and intra-sequence variations (i.e. rigid head motion), registration techniques are applied to find a displacement field T that registers each frame to a neutral reference frame, while maintaining the facial expression:

$$T = T_{inter} \circ T_{intra}. \tag{1}$$

The intra-sequence displacement field T_{intra} is modelled as a simple affine registration. The facial part of each frame in the sequence is registered to the facial part of the first frame to suppress minor head motions. This is done using a gradient descent optimization, with the squared sum of differences (SSD) of the grey level values as a distance metric.

The inter-subject displacement field T_{inter} is again modelled as an affine registration. A subset of 9 of the 20 facial points detected in the first frame that are stable (i.e., their location is mostly unaffected by facial expressions) is registered to a predefined reference set of facial points. This predefined set of reference points is taken from an expressionless image of a subject that was not used in the rest of the experiments. The displacement field T_{inter} is applied to the entire image sequence to eliminate inter-subject differences in facial shape.

The T_{intra} and T_{inter} registrations are performed separately since T_{inter} is a geometric registration of two sets of fiducial facial points, whereas T_{intra} is an appearance-based registration based on the minimization of the sum of squares of the motion-compensated image intensities. Therefore, we can not combine the two registrations. Let us also note here, that intra-sequence transforms (i.e., from a frame to the previous one)

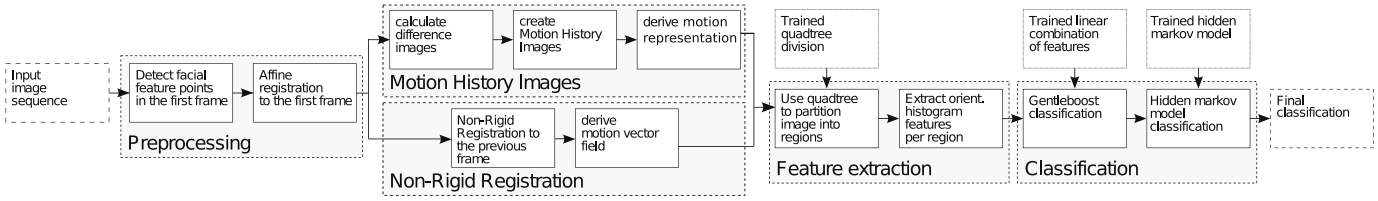


Fig. 2: Outline of the proposed method.

Fig. 3: An illustration of the rigid registration process. Also shown are the 10 facial feature points used for registration.

are in general smaller and therefore more easily estimated than the combined transform to a global reference frame. However, once estimated, T_{inter} and T_{intra} are combined and applied as a single transformation. An illustration of the two steps and the used facial points is given in Fig. 3.

3.2 Motion representation

Most existing approaches base their classification on either single frames or entire videos. Here, we use overlapping sliding windows of different sizes and classify each window in terms of depicted AUs and their temporal segments. In any given frame, each AU can be in one of four different temporal segments: neutral (inactive), onset, apex, or offset. Different AUs have different onset and offset durations. Therefore it is useful to have a flexible θ (size of temporal window) and consider several sizes. The onset of AU 45(blink), for instance, has an average duration of 2.4 frames (in the utilized datasets). On the other hand, the offset of AU 12 (smile) lasts 15.4 frames on average. A temporal window of 2 frames is well-suited to find the onset of AU 45, but it is hard to detect the onset of AU 12 using such a window. Therefore, several window sizes are tested, ranging from 2 frames to 20 frames. 96.4% of all onsets/offsets in our dataset last 20 frames or less, so this size suffices to easily capture most activations.

To represent the motion in the face due to facial expressions, two different methods of Motion History Images and Non-rigid registration using Free-form Deformations have been investigated, which will now be discussed in detail.

3.2.1 *Motion History Images*

Motion history images (MHIs) were first proposed by Davis and Bobick [8]. MHIs compress the motion over a number of frames into a single image. This is done by layering the thresholded differences between consecutive frames one over the other. In doing so, an image is obtained that gives an indication of the motion occurring in the observed time-frame.

Let t be the current frame and let θ be the temporal window size. Then, MHI_t^θ consists of the weighted layered binary difference images for each consecutive two frames $(t-\frac{\theta}{2}, t-\frac{\theta}{2}+1), \ldots, (t+\frac{\theta}{2}-2, t+\frac{\theta}{2}-1)$. A binary difference image for the pair $(t, t+1)$ is denoted with d_t and is defined as

$$d_t(x,y) = b\left(\begin{cases} 1 & |g(x,y,t) - g(x,y,t+1)| > \gamma \\ 0 & otherwise \end{cases}\right), \quad (2)$$

where $g(\cdot,\cdot,t)$ is frame t filtered by a Gaussian filter of size 2, γ is a noise threshold set to 4 (this means that two pixels must differ 4 grey levels to be classified as different), b is a binary opening filter applied to the difference image to remove remaining isolated small noise spots with an area smaller than 5 pixels. g was varied between 0 and 10, γ was varied between 1 and 20, b was varied between 0 and 20. The parameters were varied on a small set of videos and the values as used above gave the best results for recognition.

Using weighted versions of these binary difference images, the MHI is then defined as:

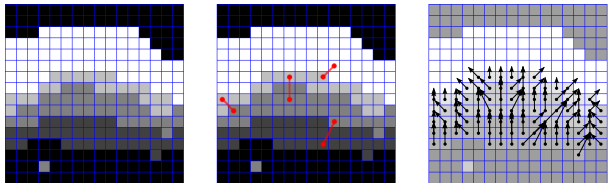
$$M_t^\theta = \frac{1}{\theta}\max_s(\{(s+1)d_{t-\frac{\theta}{2}+s} | 0 \leq s \leq \theta - 1\}). \quad (3)$$

That is, the value at each pixel of the MHI is the weight of the last difference image in the window that depicts motion, or 0 if the difference images do not show any motion.

In the original implementation by Davis [8], motion vectors are retrieved from the MHI by simply taking the Sobel gradient of the image. This will however only give motion vectors at the borders of each grey level intensity in the image. This works well in the case that the MHIs show smooth and large motion, but in our case the motion is usually shorter and over a smaller distance, leading to less smooth gradients in the image. Applying the Sobel gradient in such a case leads to a very sparse motion representation. The approach taken here is as follows. For each pixel that is not a background pixel (i.e. pixels where M_t^θ is 0 since no motion was detected), we search in its vicinity for the nearest pixel of higher intensity (without crossing through background pixels). The direction in which a brighter pixel lies (if there is one) is the direction of motion in that pixel. In the case that multiple brighter pixels are found at the same distance, the pixel closest to the centre of gravity of those pixels is chosen. This gives us a dense and informative representation of the occurrence and the direction of motion. This is illustrated in Fig. 4.

3.2.2 *Non-rigid Registration using FFDs*

This method is an adapted version of the method proposed by Rueckert et al. [29], which uses a free-form deformation (FFD) model based on b-splines. The method was originally used to register breast MR images, where the breast undergoes local shape changes as a result of breathing and patient motion.



(a) (b) (c)

Fig. 4: Illustration of the estimation of a motion vector field from an MHI. (a): Original MHI. (b): for each pixel, the closest neighbouring brighter pixel is found (without crossing background pixels). (c): This process is repeated for each pixel, resulting in the motion vector field shown here.

Let Ω_t denote the grey-level image of the face region at frame t, where $\Omega_t(x,y)$ is the grey-level intensity at pixel (x,y). Given a pixel (x,y) in frame t, let $(\hat{x},\hat{y})$ be the unknown location of its corresponding pixel in frame $t-1$. Then, the non-rigid registration method is used to estimate a motion vector field $\hat{F}_t$ between frames t and $t-1$, such that:

$$(\hat{x},\hat{y}) = (x,y) + \hat{F}_t(x,y) \tag{4}$$

To estimate $\hat{F}_t$, we select a $U \times V$ lattice Φ_t of control points with coordinates $\phi_t(u,v)$ in Ω_t, evenly spaced with spacing d. Then, non-rigid registration is used to align Φ_t with Ω_{t-1}, resulting in a displaced lattice $\hat{\Phi}_{t-1} = \Phi_t + \Phi_\delta$. Then, $\hat{F}_t$ can be derived by b-spline interpolation from Φ_δ. To estimate $\hat{\Phi}_{t-1}$, a cost function C is minimized. Rueckert et al. [29] use normalized mutual information as the image alignment criterion. However, in the 2D low-resolution case considered here, not enough sample data is available to make a good estimate of the image probability density function from the joint histograms. Therefore, we use the sum of squared differences (SSD) as the image alignment criterion, i.e. :

$$C(\hat{\Phi}_{t-1}) = \sum_{x,y} (\Omega_t(x,y) - \Omega_{t-1}(\hat{x},\hat{y}))^2 \tag{5}$$

The full algorithm for estimating $\hat{\Phi}_{t-1}$ (and therefore Φ_δ) is given in Fig. 5. We can calculate $\hat{F}_t$ using b-spline interpolation on Φ_δ.

For a pixel at location (x,y), let $\phi_t(u,v)$ be the control point with coordinate (x_0,y_0) that is the nearest control point lower and to the left of (x,y), i.e. it satisfies:

$$x_0 \le x < x_0 + d, \qquad y_0 \le y < y_0 + d \tag{6}$$

In addition, let $\phi_\delta(u,v)$ denote the vector that displaces $\phi_t(u,v)$ to $\hat{\phi}_{t-1}(u,v)$. Then, to derive the displacement for any pixel (x,y), we use a b-spline interpolation between its 16 closest neighbouring control points (see Fig. 6). This gives us the estimate of the displacement field $\hat{F}_t$

$$\hat{F}_t(x,y) = \sum_{k=0}^{3}\sum_{l=0}^{3} B_k(a) B_l(b) \phi_\delta(u+k-1, v+l-1), \tag{7}$$

where $a = x - x_0, b = y - y_0$ and B_n is the n^{th} basis function

Find the 20 facial points in the first frame of the sequence
Find T_{inter} (affine transformation to reference facial points)
Apply T_{inter} to the entire sequence
foreach frame t **do**
  Find T_{intra} (affine transformation to frame 1) and apply it
  Initialize the control point lattice $\hat{\Phi}^0_{t-1}$ as Φ^0_t
  foreach control point density d **do**
    Calculate the gradient vector of the cost function C in terms of $\hat{\Phi}^d_{t-1}$: $\nabla C = \frac{\delta C(\hat{\Phi}^d_{t-1})}{\delta \hat{\Phi}^d_{t-1}}$
    while $||\nabla C|| > \epsilon$ **do**
      Recalculate the control point positions: $\hat{\Phi}^d_{t-1} = \hat{\Phi}^d_{t-1} + \mu \frac{\nabla C}{||\nabla C||}$
      Recalculate ∇C
    end
    Increase the density d of the control point lattice
    Add points to $\hat{\Phi}^{d+1}_{t-1}$ from $\hat{\Phi}^d_{t-1}$ by b-spline interpolation
  end
  Derive Φ_δ: $\Phi_\delta = \hat{\Phi}_{t-1} - \Phi_t$
  Use b-spline interpolation to derive $\hat{F}_t$ from Φ_δ **end**

Fig. 5: The non-rigid registration algorithm. ϵ is a stopping criterion and μ is the step size in the recalculation of control point positions. The values for both are taken from [29].

of the uniform cubic b-spline, i.e.:

$$\begin{aligned} B_0(a) &= (-a^3 + 3a^2 - 3a + 1)/6, \\ B_1(a) &= (3a^3 + 6a^2 + 4)/6, \\ B_2(a) &= (-3a^3 + 3a^2 + 3a + 1)/6, \\ B_3(a) &= a^3/6. \end{aligned}$$

To speed up the process, and avoid local minima, we use a hierarchical approach in which the lattice density is being doubled at every level in the hierarchy. The coarsest lattice Φ^0_t is placed around the point $c = (c_x, c_y)$ at the intersection of the horizontal line that connects the inner eye corners, and the vertical line passing through the tip of the nose and the centre of the upper the and bottom lip. Then,

$$\Phi^0_t = \left\{ (u,v) \;\middle|\; \begin{array}{l} u \in [c_x - 2id, \ldots, c_x + 2id], \\ v \in [c_y - 2id, \ldots, c_y + 4id] \end{array} \right\} \tag{8}$$

where id is the distance between the eye pupils (i.e. Φ^0_t consists of 35 control points). New control points are iteratively added in between, until the spacing becomes $0.25id$ (approximately the size of a pupil), giving 1617 control points. This has proven sufficient to capture most movements and gives a good balance between accuracy and calculation speed.

Having estimated $\hat{F}_t$, we now have a motion vector field depicting the facial motion between frame $t-1$ and t, from which orientation histogram features can be extracted. For feature extraction, we actually consider the motion vector field sequence $\hat{F}^\theta_t$ over a sliding window of size θ around frame t.

Fig. 7 shows an example of the MHI and FFD methods. Fig. 7(a) and 7(b) show the first and last frame of the sequence. Fig. 7(c) shows the resulting MHI M^θ_t, where θ is set such as to include the entire sequence. It is quite easy for humans to recognize the face motion from the MHI. Fig. 7(d) shows the motion field sequence $\hat{F}^\theta_t$ from the FFD method applied to a rectangular grid. The face motion (Fig. 7(f)) is less clear

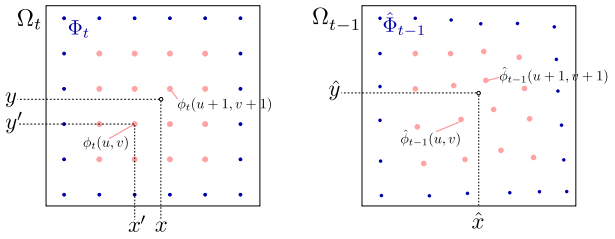


Fig. 6: Illustration of the B-spline interpolation showing an image Ω_t and the control point lattice Φ_t, as well as the estimated $\hat{\Phi}_{t-1}$ aligned with Ω_{t-1}. To estimate the new position $(\hat{x}, \hat{y})$ of the point at (x, y), only the 16 control points shown in a lighter, red colour are used.

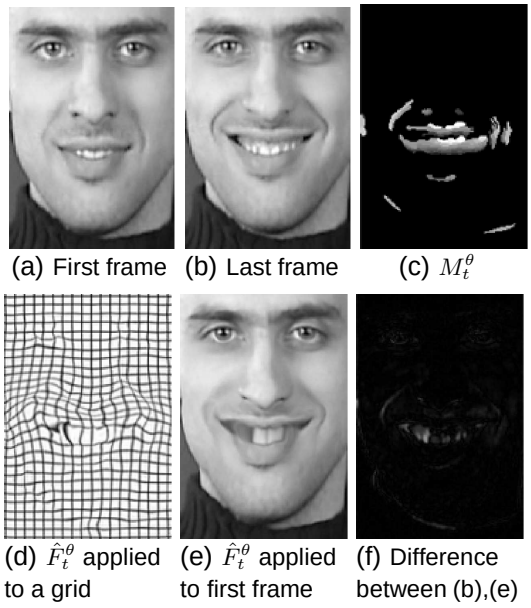


Fig. 7: Example of MHI and FFD techniques.

to the human eye from this visualization of the transform. However, when we transform the first frame by applying $\hat{F}_t^\theta$ to get an estimate of the last frame, the similarity is clear as shown in Fig. 7(e). In addition, one can see that between Fig. 7(a) and 7(b), the subject shows a slight squinting of the eyes (AU6). While this is invisible in the resulting MHI (Fig. 7(c)), it is visible in the motion field derived from FFD (Fig. 7(d)), indicating that the FFD method is more sensitive to subtle motions than the MHI method.

3.3 Feature Extraction

3.3.1 Quadtree Decomposition

In order to define the face sub-regions at which features will be extracted, we use a quadtree decomposition. Instead of dividing the face region into a uniform grid (e.g. as in [43]) or manually partitioning the face, a quadtree decomposition is used to divide the regions in a such a manner that areas showing much motion during the activation of a specific AU are divided in a large number of smaller sub-regions, while those showing little motion are divided into a small number of large sub-regions. This results in an efficient allocation of the features. We note that different features (i.e. different quadtree decompositions) are used for the analysis of different AUs.

Some AUs are very similar in appearance but differ greatly in the temporal domain. For instance, AU 43 (closed eyes) looks exactly like AU 45 (blink) but lasts significantly longer. Therefore, we also use a number of temporal regions to extract features. Let $\Theta_{a,s}$ be the collection of all sliding windows of size θ around the frames depicting a particular AU a in a particular temporal segment s in the training set. We then use a quadtree decomposition specific to each AU and the segments onset and offset on a set of projections of $\Theta_{a,s}$ to decide where to extract features to recognize the target AU and its target temporal segment.

Three projections of each window are made showing the motion magnitude, the motion over time in the horizontal direction, and the motion over time in the vertical direction:

$$P_{mag}^{\theta}(x,y) = \sum_t u(x,y,t)^2 + v(x,y,t)^2, \tag{9}$$

$$P_{tx}^{\theta}(t,x) = \sum_y u(x,y,t)^2, \tag{10}$$

$$P_{ty}^{\theta}(t,y) = \sum_x v(x,y,t)^2 \tag{11}$$

where $u(x,y,t)$ and $v(x,y,t)$ are the horizontal and vertical components of the motion vector field sequence $\hat{F}_t^\theta$. These projections are then summed over all windows in $\Theta_{a,s}$ to get the final projections used for the quadtree decomposition:

$$P_{mag}^{\Theta_{a,s}}(x,y) = \sum_{\theta \in \Theta_{a,s}} P_{mag}^{\theta}(x,y), \tag{12}$$

$$P_{tx}^{\Theta_{a,s}}(t,x) = \sum_{\theta \in \Theta_{a,s}} P_{tx}^{\theta}(t,x), \tag{13}$$

$$P_{ty}^{\Theta_{a,s}}(t,y) = \sum_{\theta \in \Theta_{a,s}} P_{ty}^{\theta}(t,y) \tag{14}$$

These three images then undergo a quadtree decomposition to determine a set of 2D regions ((x,y)-, (t,x)-, and (t,y)-regions) where features will be extracted. The defined projections show us exactly where much motion occurs for a particular AU and a particular temporal segment and where there is less motion. The quadtree decomposition algorithm is outlined in Fig. 8. The splitting threshold τ was set to 0.1, meaning a region in the quadtree will be split if the region accounts for 10% of the total motion in the frame. This gives a reasonable balance between having too large regions, so the detail is lost, and too many small regions, where the features become less effective as facial features do no longer always fall in the same region. The minimum region size σ is defined to be $0.25id$, where id is the interocular distance. In other words, the minimum region size is about the size of a pupil. Extracting features in smaller regions will not be very informative due to small variations in facial feature locations in different subjects. Some examples of motion magnitude images and the resulting quadtree decompositions are shown in Fig. 9. We can see in Fig. 9(e) that for AU46R (right eye wink) most of the features will be extracted in the eye area, where all the motion occurs.

Initialize R with a single region (the entire face region)
Define p_{total} as the summed value of all pixels in P
Define τ as the splitting threshold
Define σ as the minimum size of a region
while True **do**
  foreach region r **in** R **do**
    Calculate p_r, the summed value of all pixels in r
    if $p_r < \tau \cdot p_{total}$ **and** $size(r) > \sigma$ **then**
      Remove r from R
      Split r in 4 equally sized rectangles
      Add these to R
    end
  end
  if no region was split **then stop**
end

Fig. 8: The quadtree decomposition algorithm. τ is the threshold for splitting, σ is the minimum region size.

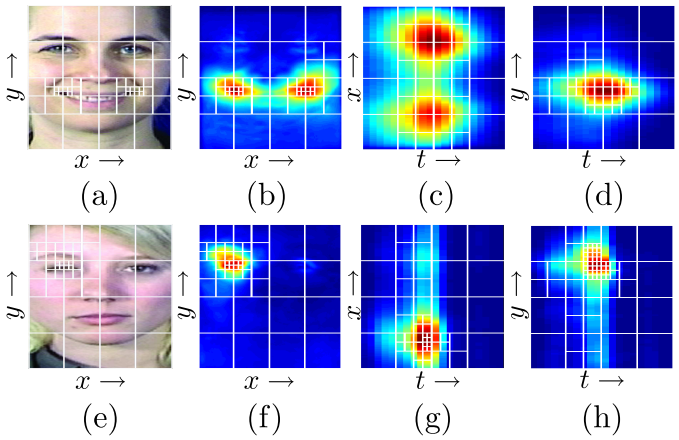


Fig. 9: Quadtree decompositions: (a,b,c,d) Onset of AU 12(smile); (e,f,g,h) Onset of AU 46R(right eye wink). Shown for each AU are example frames (a,e) and the three projections $P_{mag}^{\Theta_{a,s}}$ (b,f), $P_{tx}^{\Theta_{a,s}}$ (c,g), $P_{ty}^{\Theta_{a,s}}$ (d,h). Overlaid on each projection is the resulting quadtree decomposition.

In $\Theta_{a,s}$, some frames also show the activation of other AUs than a. Usually, the activation of other AUs does not occur frequently enough to significantly alter the decomposition. However, in some cases, AUs co-occur very frequently and the decomposition shows some of the motion of the co-occurring AU. It may then happen that some features corresponding to the co-occurring AU are then selected to classify a.

3.3.2 Features

After generating the quadtree decompositions, we extract the features for the sliding window around each frame in the dataset. We consider the $u(x, y, t)$ and $v(x, y, t)$ components from $\hat{F}_t^\theta$ in the sub-regions determined by the quadtree decomposition of $P_{mag}^{\Theta_{a,s}}(x, y)$. In each sub-region 11 features are extracted from the components: an orientation histogram of 8 directions, the divergence, the curl, and the motion magnitude.

For the temporal regions determined by the decompositions of $P_{tx}^{\Theta_{a,s}}(t, x)$ and $P_{ty}^{\Theta_{a,s}}(t, y)$, we first determine the projections $P_{tx}^{\theta}(t, x)$ and $P_{ty}^{\theta}(t, y)$ for the test frame in question. For each sub-region in the projections, we extract 3 features: the average absolute motion, the average amount of positive (i.e. left, upward) motion and the average amount of negative (i.e. right, downward) motion.

AU	1	5	9	12	16	24	27
onset original	2013	2013	1551	1551	1551	1650	1386
onset selected	67	67	34	47	19	87	12
offset original	1452	1815	1551	1683	1551	1749	1683
offset selected	90	85	76	86	34	86	73

TABLE 1: Original number of features and number of features selected by GentleBoost per AU when trained on the entire MMI dataset with a window-size of 20 frames.

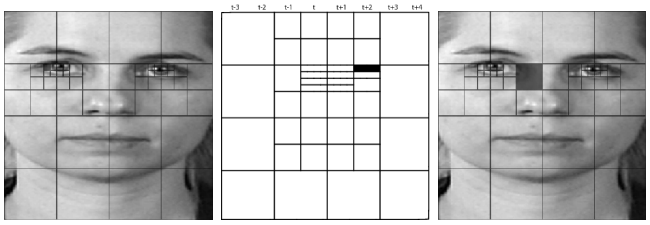
3.4 Classification

We use the GentleBoost algorithm [12] for feature selection and classification. Advantages of GentleBoost over AdaBoost are that it converges faster and is more reliable when stability is an issue [12]. For each AU and each temporal segment characterised by motion (i.e. onset, offset), we train a dedicated one-vs-all GentleBoost classifier. Since our dataset is rather unbalanced (over 95% of the frames in the database depict expressionless faces), we initialize the weights such that both the positive and the negative classes carry equal weight. This prevents that all frames are classified as neutral. The GentleBoost algorithm is used to select a linear combination of features one at a time until the classification no longer improves by adding more features. This gives a reasonable balance between speed and complexity. The number of features selected for each classifier range between 19 and 93, with an average of 74 features selected. Table 1 gives an overview of the number of selected features for several AUs.

The first three selected features for some of the classifiers are shown in Figures 10-11. In the images, for each feature selected from the $P_{mag}^{\Theta_{a,s}}$-projection, a neutral face image is overlaid to indicate the location of the region. The selected features correspond reasonably well to the intuitively interesting features/regions for each AU. The $P_{mag}^{\Theta_{a,s}}$-projection is the most important (and most often selected) projection since most information is available in the spatial domain. This is also the reason why the problem of facial expression recognition can be solved (to a certain extent) using static images (e.g. [26]). However, for some AUs, the information in the spatial magnitude projection is insufficient to distinguish them from other AUs. One example is AU 43 (closed eyes), which only differs from AU 45 (blink) in the temporal domain. Since AU 45 is much more common, an AU 43 detector that does not take the temporal domain into account would detect many false positives. Fig. 11 shows that a temporal feature is the second most important one in the detection of the onset of AU 43. The feature in question measures the amount of upward motion in the eyelid area for the next 2 frames. If the depicted AU were AU 45, then the next 2 frames after any of the onset frames should show upward motion as the eye would be be opening again. In AU 43 however, the next 2 frames after any of the onset frames will show no motion as the eyes will still be closed. Thus, the absence of upward motion in this area in a period of 2 frames after an onset frame is a very good way to

(a) $P^{\Theta_{1,8}}_{mag}$: divergence (b) $P^{\Theta_{1,8}}_{mag}$: divergence (c) $P^{\Theta_{1,8}}_{mag}$: divergence

Fig. 10: First three selected features for onset of AU 1 (inner brow raiser), window size 8, superimposed on a neutral frame.



(a) $P^{\Theta_{43,8}}_{mag}$: divergence (b) $P^{\Theta_{43,8}}_{ty}$: no upward motion (c) $P^{\Theta_{43,8}}_{mag}$: divergence

Fig. 11: First three selected features for onset of AU 43 (closed eyes), window size 8, superimposed on a neutral frame. (b) depicts the absence of upwards motion in shown y-area of frame $t+2$.

tell apart AU 43 from AU 45 onset segments.

Each onset/offset GentleBoost classifier returns a single number per frame indicating the confidence that that frame depicts the target AU and the target temporal segment. In order to combine the onset/offset GentleBoost classifiers into one AU recognizer, a continuous HMM is used. The motivation for using an HMM is to use the knowledge that we can derive from our training set about the prior probabilities of each temporal segment of an AU and its duration (represented in the HMM's transition matrix). Hence, an HMM is trained for the classification of each AU.

HMMs are defined by $\lambda = \{\Lambda, B, \Pi\}$, where Λ is the transition matrix, B is the emission matrix and Π is the initial state probability distribution. These are all estimated from the training set, where the outputs of the onset- and offset-GentleBoost classifiers are used to calculate the emission matrix B for the HMM by fitting a Gaussian to the values of both outputs in any temporal state. Then, the probability for each state can be calculated given the output of the GentleBoost classifiers in a particular frame.

The HMM has four states, one corresponding to each of the temporal segments. The initial probabilities Π show that the sequences in our dataset usually start in the neutral segment (i.e. no AU is depicted), but on rare occasions the AU is already in one of the other states. Based on the initial probabilities Π, the transition probabilities Λ and emission probability matrix B, the HMM decides the mostly likely path through the temporal segment states for the input image sequence, using the standard Viterbi algorithm. This results in the classification of the temporal segment for each frame in the tested image sequence.

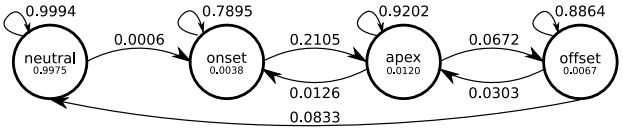


Fig. 12: The states and transition probabilities for an HMM trained on AU 1. Initial probabilities are denoted below the state names. Transitions with probability 0 are not shown.

The HMM facilitates a degree of temporal filtering. For instance, given that the input data temporal resolution is 25 fps and given the facial anatomy rules, it is practically impossible to have an apex followed by a neutral phase and this is reflected in the transition probabilities Λ. Also, the HMM tends to smooth out the results of the GentleBoost classifiers (for instance, short incorrect detections are usually filtered out). However, it only captures the temporal dynamics to a limited degree, since it operates under the Markov assumption that a signal value at time t is only dependent on the signal value at time $t-1$. For example, the HMM does not explicitly prevent onsets that last only one frame (even though in most AUs, the minimum onset duration is much longer). Yet it does model these dynamics implicitly through its use of transition probabilities between the states.

An example of the learned transition probabilities Λ for one HMM, trained to recognize AU 1, is given in Fig. 12. The transition probabilities say something about the state duration. For instance, the transition probability for *neutral* $\rightarrow$ *neutral* is very high, since the duration of a neutral state is usually very long (it is as long as the video itself when the video does not contain the target AU). The normal sequence of states is *neutral* $\rightarrow$ *onset* $\rightarrow$ *apex* $\rightarrow$ *offset* $\rightarrow$ *neutral*. However, the transition probabilities show that, although highly unlikely, transitions *apex* $\rightarrow$ *onset* or *offset* $\rightarrow$ *apex* do occur. This is typical for spontaneously displayed facial expressions which are characterized by multiple apexes [11, 23]. As both utilized datasets, the MMI and the Cohn-Kanade dataset, contain recordings of acted (rather than spontaneously displayed) facial expressions, occurrence of multiple apexes is rare and unlikely. In the SAL spontaneous expression dataset on the other hand, multiple apexes occur quite frequently. However, especially in the MMI dataset and especially by brow actions (AU1, AU2), smiles (AU12), and parting of the lips (AU25), some recordings seem to be capturing spontaneous (unconsciously displayed) rather than purely acted expressions.

4 Experiments

4.1 Datasets

The first dataset consists of 264 image sequences taken from the MMI facial expression database [27] (www.mmifacedb.com). To the best of our knowledge, this data is the largest freely available dataset of facial behaviour recordings. Each image sequence used in this study depicts a (near-)frontal view of a face showing one or more AUs. The image sequences are chosen such that all AUs under

consideration are present in at least ten of the sequences and distributed over 15 subjects. The image sequences last on average 3.4 seconds and were all manually coded for the presence of AUs. Ten-fold cross-validation was used, with the folds divided such that each fold contains at least one example of each AU. Temporal window sizes ranging from 4 to 20 frames were all tested independently and the window size that yielded the best result was chosen.

To test the generalization performance of the system, we have also evaluated the proposed FFD-based method on the Cohn-Kanade (CK) dataset [15], arguably the most widely used dataset in the field. We only tested the system on those AUs for which more than ten examples existed in the CK dataset. This resulted in examples of 18 AUs shown in 143 sequences in total. The original CK dataset only has event coding for the AUs (stating only whether an AU occurs in the sequence, not a frame-by-frame temporal segment coding). Here, we have used frame-by-frame annotations provided by Valstar& Pantic [34] based on the given event coding.

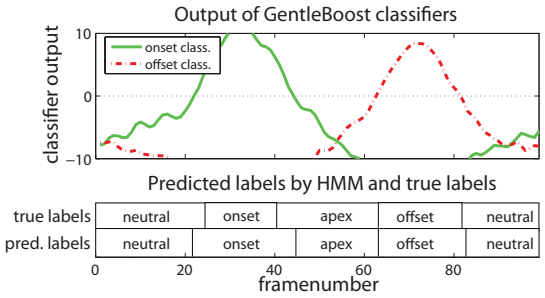
Finally, we also tested the method on the SAL (Sensitive Artificial Listener) dataset containing displays of spontaneous expressions [9]. The expressions were elicited in human-computer conversations through a 'Sensitive Artificial Listener' interface. Subjects converse with one of four avatars, each having its own personality. The idea is for subjects to unintentionally and spontaneously mirror the emotional states of the avatars. 10 subjects were recorded for around 20 minutes each. The speech sections were removed from the data, leaving 77 sequences that depict spontaneous facial expressions. For 4 subjects, the data has been FACS-coded on a frame-by-frame basis, for the other 6 subjects only event coding exists. Since our method requires frame-by-frame annotations to train the classifiers, we used data of 4 subjects for training and we tested on the remaining 6 subjects. We only tested our method on the 10 AUs for which there were at least 5 training examples.

4.2 Results

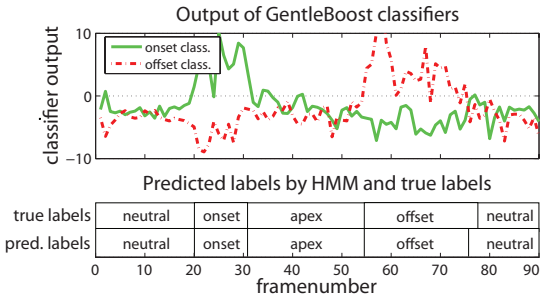
Fig. 13 shows two typical results for AU 27 (mouth stretch). As can be seen in Fig. 13(a), the GentleBoost classifiers yield good results and the resulting labelling is almost perfect for $\theta = 20$. For $\theta = 2$, the GentleBoost classifiers yield less smooth results (Fig. 13(b)). Even so, the HMM filters out the jitter very effectively.

4.2.1 Event Coding

Table 2 gives the results for all AUs tested with the MHI and the FFD technique on the MMI dataset (per AU, the window width θ that gave the highest F_1-score is mentioned). The F_1-measure is a weighted mean of the precision and recall measures. In the manual labelling of the dataset, AU 46 (wink) has been split up into 46L and 46R, since the appearance differs greatly depending on which eye is used to wink. Similarly, AU 28 (lip suck) is scored when both lips are sucked into the mouth, and AU 28B and AU 28T are scored when only the lower or the upper lip is sucked in. This gives us a total of 30 classes, based on the 27 AUs defined in FACS. As can be seen from Table 2, both techniques



(a) AU 27, $\theta = 20$



(b) AU 27, $\theta = 2$

Fig. 13: Example classification results. Top: The output of the GentleBoost-classifiers. Bottom: The true and estimated frame labels (as predicted by the HMM). θ is the used temporal window size.

	Results FFD method					Results MHI method				
AU	θ	CR	RC	PR	F_1	θ	CR	RC	PR	F_1
1	20	97.7	61.5	88.9	**72.7**	20	93.9	53.9	41.2	**46.7**
2	20	97.7	66.7	80.0	**72.7**	20	96.2	50.0	60.0	**54.6**
4	20	91.3	74.3	65.0	**69.3**	20	76.1	91.2	34.1	**49.6**
5	20	93.6	66.7	38.1	**48.5**	12	93.6	27.3	25.0	**26.1**
6	20	96.2	82.4	66.7	**73.7**	20	93.6	76.9	41.7	**54.1**
7	8	92.1	54.6	27.3	**36.4**	8	86.0	45.5	13.9	**21.3**
9	20	97.0	81.8	60.0	**69.2**	20	93.6	70.0	33.3	**45.2**
10	20	97.4	78.6	73.3	**75.9**	20	95.8	42.9	66.7	**52.2**
11	12	94.7	77.8	58.3	**66.7**	16	89.0	33.3	26.1	**29.3**
12	20	93.6	82.4	50.0	**62.2**	20	80.3	100	24.6	**39.5**
13	12	95.5	90.0	45.0	**60.0**	12	87.9	20.0	7.7	**11.1**
14	16	91.3	75.0	38.7	**51.1**	16	91.7	68.8	39.3	**50.0**
15	8	94.7	75.0	45.0	**56.3**	20	95.1	25.0	42.9	**31.6**
16	16	97.0	85.7	66.7	**75.0**	16	95.8	57.1	61.5	**59.3**
17	16	83.7	75.3	77.8	**76.5**	20	74.2	86.7	58.2	**69.6**
18	16	91.7	63.6	50.0	**56.0**	12	84.9	47.8	28.2	**35.5**
20	20	95.1	45.5	41.7	**43.5**	20	91.3	36.4	20.0	**25.8**
22	12	93.2	72.7	34.8	**47.1**	20	94.3	36.4	33.3	**34.8**
23	16	92.4	58.3	31.8	**41.2**	16	91.3	8.3	7.7	**8.0**
24	16	89.4	61.1	34.4	**44.0**	16	89.0	20.0	15.0	**17.1**
25	8	90.5	92.0	78.4	**84.7**	20	71.6	86.7	50.0	**63.4**
26	20	95.5	81.8	81.8	**81.8**	20	82.2	61.3	35.2	**44.7**
27	20	99.6	100	92.9	**96.3**	20	95.8	100	54.2	**70.3**
28	16	93.6	92.9	44.8	**60.5**	20	88.6	42.9	21.4	**28.6**
28B	16	95.5	72.7	47.1	**57.1**	12	92.8	36.4	25.0	**29.6**
28T	12	92.4	80.0	30.8	**44.4**	16	84.5	50.0	12.2	**19.6**
43	20	95.1	60.0	56.3	**58.1**	20	86.4	20.0	11.1	**14.3**
45	8	93.6	90.8	93.4	**92.1**	4	85.6	96.3	75.4	**84.6**
46L	8	99.2	90.9	90.9	**90.9**	8	97.0	54.6	66.7	**60.0**
46R	8	99.2	81.8	100	**90.0**	12	97.0	27.3	100	**42.9**
avg	-	94.3	75.7	59.7	**65.1**	-	89.2	52.4	37.7	**40.6**

AU = Action Unit, θ = Window Size, CR = Classification Rate
RC = Recall Rate, PR = Precision Rate, F_1 = F_1-measure

TABLE 2: Results for 27 AUs (30 classes) on 264 sequences from the MMI dataset for the MHI and the FFD method.

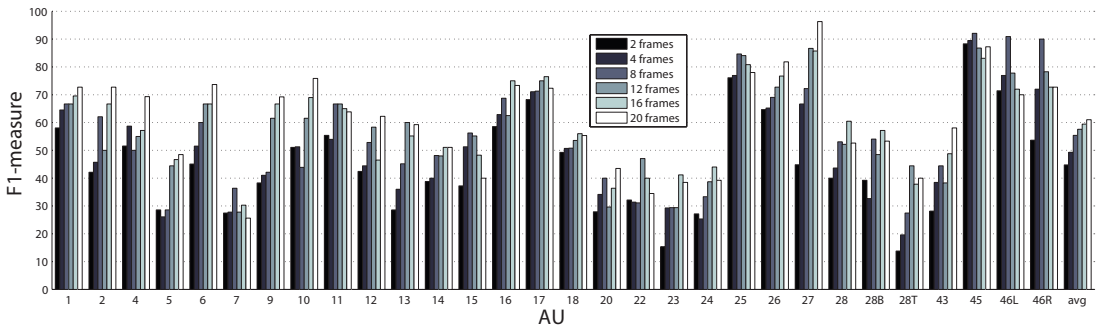


Fig. 14: F_1-measure per AU for different window sizes for the FFD method.

have difficulties with subtle AUs (i.e. 5 (upper lid raiser), 7 (eye squint), 23 (lip tightener)). These problems possibly stem from the method of extracting motion statistics over larger regions. If the regions are too large, these subtleties are easily lost (however, having the regions too small generates errors relating to the rigid registration and inter-subject differences). Possibly, geometric approaches are better equipped to handle these AUs (e.g. AU5, AU7), since their activation is clearly observable from displacements of facial fiducial points and no averaging of the motion over regions is needed.

It is clear that, overall, the FFD technique produces superior results to those obtained for the MHI-based approach. Therefore, in the remainder of this work, only the FFD-based approach is investigated further. One reason for the inferior performance of the MHI-based approach is that only intensity differences above the noise threshold are registered in the MHI. For instance, if the mouth corner moves (e.g. in AU12), only the movement of the corner of the mouth is registered in the related MHI. More subtle and smoother motion of the skin (e.g., on the cheeks) is not registered in the related MHI (see Fig. 7). In the FFD method however, we will see the entire cheek deform as a result. Also, in MHIs earlier movements can obscure later movements (e.g. in AU 28) and fast movements can show up as disconnected regions that do not produce motion vectors (e.g. in AU 27).

In general, the F_1-measure is reasonably high for most AUs when the FFD technique is applied, but there is still room for improvement. In particular, there are many false positives. Most of these occur in AUs that have a similar appearance. The AUs performing below 50% are AUs 5 (upper lid raiser), 7 (eye squint), 20 (lip stretcher), 22 (lip funneller), 23 (lip tightener) and 28T (upper lip inward suck). For most of these AUs, the reasons for the inaccurate performance lie in the confusion of the target AU with other AUs. For instance, the onset of AU7 (eye squint) is often confused with the onset of AU45 (blink), the offset of AU5 is very similar to the onset of AU45 (and vice versa), and AUs 20, 23, 24 and 28T are often confused with each other since they all involve downward movement of the upper lip.

Another cause of some false positives is a failure of the affine registration meant to stabilize the face throughout the sequence. Out-of-image-plane head motions, for instance, if not handled well, result in some classifiers classifying rigid face motions as non-rigid AU activations. We partially address this issue for spontaneous expressions in Section 4.2.3 by incorporating the results of a facial point tracker in the rigid registration process. However, we should note that for very large out-of-plane rotations, affine registration is not sufficient. The use of 3D models seems a promising direction. However, they require the construction of a 3D model that might be difficult to obtain from monocular image sequences.

Though most AUs perform best with the largest window size tested, it is clear from the results that AUs with shorter durations such as AU 45 benefit from a smaller window size.

Fig. 14 shows the results for all AU classifiers for all tested window widths for the FFD technique. Overall, we see that the F_1-measure improves as the temporal window increases. Exceptions include AUs with particularly short durations, such as 7 (eye squint), 45 (blink), 46L (left eye wink), and 46R (right eye wink).

4.2.2 Temporal Analysis

We were also interested in the timing of the temporal segment detections with respect to the timing delimited by the ground truth. This test was run using the optimal window widths as summarized in Table 2. Only sequences that were correctly classified in terms of AUs were considered in this test. Four different temporal segment transitions can be detected, *neutral* $\rightarrow$ *onset*, *onset* $\rightarrow$ *apex*, *apex* $\rightarrow$ *offset*, and *offset* $\rightarrow$ *neutral*. Fig. 15 shows the average absolute frame deviations per AU and temporal segment transition. The overall average deviation is 2.46 frames. 44.12% of the detections are early and 38.18% are late. The most likely cause of late detection is that most AUs start and end in a very subtle manner, visible to the human eye but not sufficiently pronounced to be detected by the system. Early detections usually occur when a larger temporal window width is used, where the AU's segment in question is already visible in the later frames of the window, but it is not actually occurring at the frame under consideration (this can also be seen in Fig. 13a). In general, AUs of shorter duration also show smaller deviations. Also, the transitions that score badly are usually subtle ones. The high deviations for *apex* $\rightarrow$ *offset* in AUs 6 (cheek raiser and lid compressor) and 7 (eye squint) can be explained by considering that these transitions are first only slightly visible in the higher cheek region before becoming apparent in the motion of the eyelids. Since the eyelid motion is much clearer, our method targets that motion and misses the cheek raising in the start of the transition.

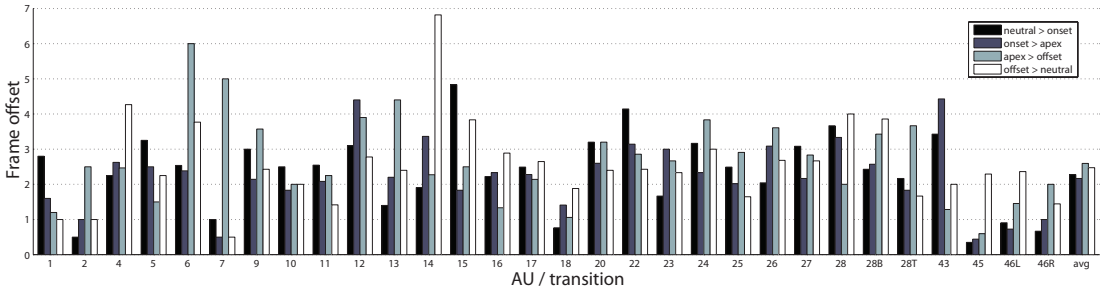


Fig. 15: Average detection offsets per AU and temporal segment transition.

Similarly, the *offset* → *neutral* transition in AU 14 (mouth corner dimpler) has almost all of the motion in the first few frames and then continues very slowly and subtly. Our method picks up only the first few frames of this transition.

Another way to look at the temporal analysis results is to analyse them per window size and transition type. Fig. 16 illustrates that. It shows the proportion of early, timely, and late detections for all correctly detected transitions per window size. It also shows the mean absolute frame offset per transition and per window size (this is depicted by the narrow bar, placed on the right side of each of the main bars in the graph). Interestingly, for the *neutral* → *onset* and *apex* → *offset* transitions the most accurate results are obtained for the lowest window size and the results deteriorate as the window size increases. For the other two transitions, the lower window sizes are actually less accurate and the best results are obtained at window sizes 8 and 12. This behaviour might be explained by a few factors. Firstly, most motion occurs in the beginning of the onset and offset segments, with the endings of those segments containing slower, more subtle motions. Hence, the transitions indicating the end of motion (*onset* → *apex* and *offset* → *neutral*) are detected early since the subtle motion at the end of the onset and offset segments remains undetected by the system. The transitions indicating the start of motion (*neutral* → *onset* and *apex* → *offset*) are quite unlikely to be early, simply because there is no prior motion which could be classified as the transition in question. The results change as the window size increases. This is due to the smoothing effect discussed earlier, due to which the start of motion is detected earlier and the end of motion is detected later.

4.2.3 Spontaneous Expressions

We performed tests on the SAL dataset, containing 77 sequences of spontaneous expressions, mostly smiles and related expressions. We tested for the 10 AUs that occurred 5 or more times. We trained on the sequences of 4 of the 10 subjects, that were annotated frame-by-frame for AUs, and tested on the data of the other 6 subjects, that were annotated per sequence.

The dataset contains relatively large head motions and moderate out-of-plane rotations. We note that in the datasets used in this paper all facial fiducial points were visible at all times. If that is not the case, one could train a different set of classifiers for each facial viewpoint.

The results for the SAL dataset are given in Table 3. The obtained classification rate is 80.2%, which is lower than the results on the posed data sets (89.8% on CK and 94.3% on MMI). However, we achieve a satisfactory average F_1-score of 75.5%, which is in fact higher than for the MMI (65.1%) and CK (72.1%) datasets. The worst performance is reported for AUs 2, 7, and 10. AUs 2 and 10 are much exaggerated in posed expressions and therefore harder to detect in subtle spontaneous depictions. AU 7 is here also often confused with AU 45, just as in the MMI dataset. The best performing AUs are 12, 25, and 6. In fact, these AUs perform much better than in the MMI dataset. This can be explained by the fact that many more training samples were available here, indicating that more training examples can greatly benefit the performance. In addition, these AUs also occur more frequently in the test set than in the MMI case, making the test set less unbalanced compared to the other datasets. We note that here the selected window sizes are much shorter than for the MMI dataset. A possible explanation for this is that spontaneous expressions are generally less smooth and depict multiple apexes interleaved with onset and offset segments. As a result, each segment occurs for a shorter time-period.

AU	θ	NT	CR	RC	PR	F_1
1	12	8	92.86	60.00	75.00	66.67
2	20	10	88.10	57.14	66.67	61.54
6	4	28	85.71	85.71	96.77	90.91
7	4	7	57.14	42.86	60.00	50.00
10	8	13	66.67	80.00	52.17	63.16
12	2	35	95.24	94.87	100.00	97.37
23	12	6	83.33	91.67	64.71	75.86
25	2	33	92.86	92.86	100.00	96.30
26	4	18	76.19	76.32	96.67	85.29
45	16	17	64.29	53.33	94.12	68.09
avg	-	-	80.24	73.48	80.61	75.52

AU = Action Unit, F_1 = F_1-score
NT = No. of training examples, CR = Classification Rate
RC = Recall Rate, PR = Precision Rate, θ = Window Size

TABLE 3: Results for testing the system for 10 AUs on 77 sequences from the SAL dataset for the FFD method.

4.2.4 Generalization Performance

To test the robustness and generalization ability of the proposed FFD method, we performed a smaller test on the Cohn-Kanade (CK) dataset [15]. We only tested on those AUs for which at least ten examples exist in the dataset (18 AUs in 143 sequences). The 10-fold cross-validation results are shown in Table 4. As a reference, the F_1-scores for the MMI dataset are also repeated. The results achieved for the CK dataset are on average similar to those for the MMI dataset. AUs 2, 5, 12, 15,

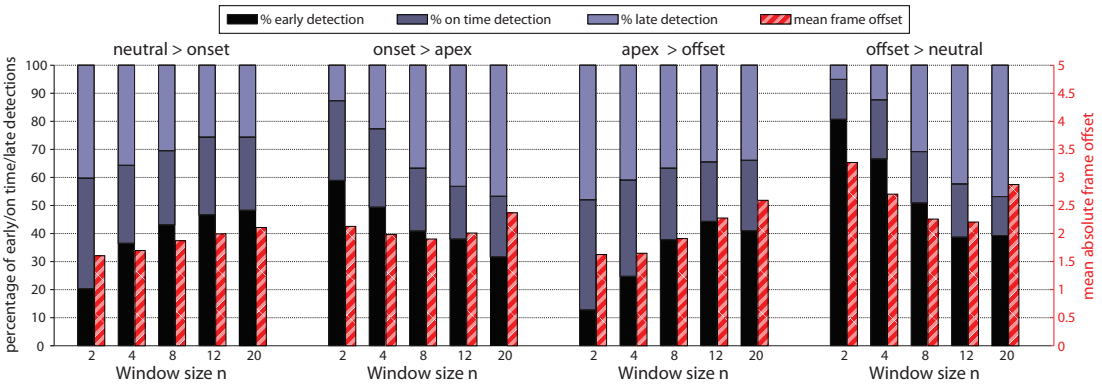


Fig. 16: Percentages of early/on time/late detection per transition and window size. Also shows average frame offset.

20, 24 and 25 perform much better in the CK dataset. Possible explanations for the inferior performance of AU 10, 11, 14 and 45 lie in the differences in ground truth labelling and the absence of offset segments in the CK dataset. The two datasets were labelled in different ways. More specifically, in the CK database, trace activations (FACS intensity A) were also coded, whereas in the MMI dataset only AUs of FACS intensity B and higher were considered. Trace activations (especially in AU 10, 11 and 14) involve very subtle changes in the facial skin appearance, that remain undetected by our method.

Another difference between the results is that for the CK dataset, lower window sizes are selected than for the MMI dataset. Since each sequence in the CK dataset ends at the apex of the expression with the offset segments cut off, no GentleBoost classifiers could be trained for the detection of offsets and the HMM classification relies solely on the onset detections. Since the duration of onsets is generally shorter than offsets, shorter window sizes tend to be selected. The absence of offset phases, especially for fast AUs like AU 45, in which onset phases can often not be captured in more than 1-2 frames and the detection relies heavily on the detection of offset phases, explains the inferior performance for such AUs. A possible explanation for better performance for AU 2, 5, 12 and 15 lies in the intensity of these expressions present in the CK dataset. More specifically, facial expression displays constituting the CK dataset are shorter and more exaggerated than it is the case with data from the MMI dataset. The better performance for AUs 24 and 25 can be explained by the greater number of examples present in the CK dataset.

We compare our results to those reported earlier by Valstar & Pantic [34], the only other authors that addressed the problem of AU temporal segments recognition. Valstar & Pantic use 153 sequences from the CK dataset, where we use 143. Their geometric-feature-based approach gives on average very similar results. Interestingly, on this dataset, the results of Valstar & Pantic are much better for AUs 4 and 7 (the related facial displays are characterized by large morphological changes which can easily be detected based on facial point displacements) and the results obtained by the FFD-based method are much better for AUs 15, 20 and 24 (which activations involve distinct changes in skin texture without large displacements of facial fiducial points). Also, the method of Valstar & Pantic is unable to deal at all with AUs 11 (nasolabial furrow deepener), 14 (mouth corner dimpler) and 17 (chin raiser), the activation of which is only apparent from changes in skin texture and cannot be uniquely detected from displacements of facial fiducial points only [26, 23].

AU	θ	CR	RC	PR	F_1	F_1^V	F_1^{MMI}
1	2	88.81	86.89	86.89	86.89	**87.6**	72.73
2	4	94.41	92.31	87.80	90.00	**94.0**	72.73
4	20	74.83	85.96	63.64	73.13	**87.4**	69.33
5	2	92.31	75.86	84.62	**80.00**	78.3	48.48
6	16	94.41	84.21	76.19	80.00	**88.0**	73.68
7	16	71.33	72.00	34.62	46.75	**76.9**	36.36
9	8	93.01	89.47	68.00	**77.27**	76.4	69.23
10	16	89.51	46.67	50.00	48.28	**50.0**	75.86
11	4	88.81	50.00	37.50	**42.86**	—	66.67
12	8	95.10	90.00	78.26	83.72	**92.1**	62.22
14	8	93.01	33.33	42.86	**37.50**	—	51.06
15	8	92.31	68.42	72.22	**70.27**	30.0	56.25
17	4	83.92	72.55	80.43	**76.29**	—	76.50
20	20	90.91	73.53	86.21	**79.37**	60.0	43.48
24	4	90.21	70.59	57.14	**63.16**	14.3	44.00
25	2	95.10	92.68	98.70	**95.60**	95.3	84.66
27	8	95.80	95.45	80.77	87.50	**89.3**	96.30
45	2	92.31	81.48	78.57	**80.00**	—	92.09

Averages	CR	RC	PR	F_1
Average our method, 18 AUs	89.78	75.63	70.25	72.14
Average our method, 14 AUs	89.86	80.29	73.21	75.85
Average [34], 14 AUs,	90.3	73.3	79.8	72.83

AU = Action Unit, θ = Window Size, CR = Classification Rate
RC = Recall Rate, PR = Precision Rate, F_1 = F_1-measure CK dataset
F_1^V = F_1 Valstar & Pantic [34], F_1^{MMI} = F_1 on MMI dataset

TABLE 4: Results for testing the system for 18 AUs on 143 sequences of the CK dataset.

Test	CR	RC	PR	F_1
Trained on MMI, tested on CK	**82.52**	**55.17**	**65.95**	**56.13**
Trained on MMI, tested on MMI	93.52	76.02	58.79	65.40
Trained on CK, tested on CK	89.78	75.63	70.25	72.14

CR=Classification Rate, RC=Recall Rate, PR=Precision Rate, F_1=F_1-measure

TABLE 5: Results for cross database testing, 18 AUs.

A cross-database test was also performed with the MMI and CK dataset. Average results are shown in table 5. The tests were run on those AUs available in both datasets using a temporal window size of 20 frames. The average result is slightly lower than the result for training and testing on the MMI dataset, but this is to be expected given the different coding styles and other differences between the two datasets.

Authors	†	features	classification
Bartlett et al. 2005 [3]	a,f	Gabor filters	AdaBoost+SVM
Bartlett et al. 2006 [4]	a,f	Gabor filters	AdaBoost+SVM
Chang 2006 [5]	a,f	manifold embed.	Bayesian
Whitehill & Omlin 2006[39]	a,f	Haar wavelets	AdaBoost
Littlewort et al. 2006 [18]	a,f	Gabor filters	AdaBoost+SVM
Lucey et al. 2007 [20]	a,f	AAM	SVM
Valstar & Pantic 2004 [36]	a,t	MHIs	kNN/rule-based
Pantic & Patras 2005 [22]	g,t	tracked face points	temporal rule-base
Valstar & Pantic 2006 [34]	g,t	tracked face points	AdaBoost+SVM
Valstar & Pantic 2007 [35]	g,t	tracked face points	AdaBoost+SVM
Tong et al. 2007 [33]	a,t	Gabor filters	AdaBoost+DBN
This work	a,t	FFD	GentleBoost+HMM

†: geometric/appearance-based(g/a), temporal-/frame-based(t/f)

TABLE 6: Comparison of AU recognition methods.

Authors	AU	NS	CR	F_1	FA	Hit	FRR
CK dataset, Image-based works							
Bartlett '05 [3]	17	313i	94.8	-	3.9	60.2	-
Bartlett '06 [4]	20	2568i	90.9	-	8.2	80.1	-
Chang '06 [5]	23	258i	89.4	-	-	-	-
Whitehill '06 [39]	11	580i	92.4	-	-	-	-
Littlewort '06 [18]	7	313i	92.9	-	-	-	-
Lucey '07 [20]	15	(?)i	95.5	-	16.7	-	1.9
CK dataset, Sequence-based works							
Valstar '04 [36]	10	344s	68	-	32.0	-	-
Pantic '05 [22]	21	90s	93.3	-	-	-	-
Valstar '06 [34]	15	(?)s	90.2	72.9	-	-	-
Tong '07 [33]	14	(?)s	93.3	-	5.5	86.3	-
This work	18	143s	89.8	72.1	6.4	75.6	29.1
This work, 15 best AUs	15	143s	92.5	72.5	4.8	73.8	26.1
MMI dataset, Image-based works							
Chang '06 [5]	29	584i	91.9	-			
MMI dataset, Sequence-based works							
Valstar '04 [36]	22	253s	61	-			
Pantic '05 [22]	9	45s	86.7	-			
Valstar '07 [35]	23	196s	-	66.0			
This work	27	264s	94.3	65.1			

AU = No. of AUs recognized, NS = number of sequences/frames (s/f) used
CR = Classification Rate, F_1 = F_1-measure, FA = False Alarm/Accept Rate
Hit = Hit Rate, FRR=False Rejection Rate

TABLE 7: Comparison of results on CK and MMI dataset.

4.2.5 Comparison to earlier work

We compared our method to earlier works that reported results on either the CK or the MMI dataset. Table 7 gives an overview of these works. It is interesting to note that most works are image-based, which means they derive the classification per frame independently and do not take temporal information into consideration. Additionally, it means that the results reported for those works are found using manually selected "peak" frames, that is, frames showing the AU in question at maximum intensity. In contrast, sequence-based approaches take the whole sequence into account without prior information as to the location of the peak intensity.

Table 6 shows results reported previously on the CK and MMI datasets. While the classification rate (the percentage of correctly classified frames/sequences) is the most commonly reported measure, it is also the one that is the least informative. Especially in cases where the dataset is highly unbalanced, it can be misleading. For example, in our subset of the CK dataset, the percentage of true positive sequences is below 10% for most AUs. This means that it is possible to report a 90% classification rate by simply classifying every sequence as negative. Therefore, we report the F_1-measure, which gives a better understanding of the quality of the classifier. Our results in terms of the classification rate on the CK dataset are largely comparable to those reported in the other works, 89.8% vs. 90.2%, 93.3%. For the MMI dataset, we outperform the other works. The main reason for the worse comparative performance on the CK dataset is probably the absence of offset segments. In contrast, both the MMI and SAL dataset contain the offset segments, which can greatly help validate the occurrence of AUs in our HMM classification scheme.

5 Conclusion and Future Work

In this work we have proposed a method based on non-rigid registration using free form deformations to model dynamics of facial texture in near-frontal-view face image sequences for the purposes of automatic frame-by-frame recognition of AUs and their temporal dynamics. To the best of our knowledge, this is the first appearance-based approach to facial expression recognition that can detect all AUs and their temporal segments. We have compared this approach to an extended version of the previously proposed approach based on Motion History Images. The FFD-based approach was shown to be far superior. On average, it achieved an F_1-score of 65% on the MMI facial expression database, 72% on the Cohn-Kanade database and 76% on the SAL dataset (containing spontaneous expressions). For each correctly detected temporal segment transition, the mean of the offset between the actual and the predicted time of its occurrence is 2.46 frames. We have compared the proposed FFD-based method to that of Valstar & Pantic [34, 35], which is the only other existing approach to recognition of AUs and their temporal segments in frontal view face images (using a geometric-feature-based approach rather than an appearance-based approach). Comparable results have been achieved for the CK facial expression database. The two approaches seem to complement each other, with some AUs being better detected with one approach and some AUs being better detected with the other approach. This is in accordance to the previously reported findings suggesting that combining the appearance- and geometric-feature-based approaches to facial expression analysis will result in an increased performance [31, 21]. Attempting to fuse the two approaches therefore seems a natural extension of this work.

References

[1] E. Aarts. Ambient intelligence drives open innovation. *ACM Interactions*, 12(4):66–68, 2005.

[2] K. Anderson and P. McOwan. A real-time automated system for recognition of human facial expressions. *IEEE Trans. Systems, Man and Cybernetics*, 36(1):96–105, 2006.

[3] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *IEEE Conf. Comp. Vision and Pattern Recognition*, pages 568–573, 2005.

[4] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 223–230, 2006.

[5] Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold-based analysis of facial expression. *Journal for Image and Vision Computing*, 24(6):605–614, 2006.

[6] D. Chetverikov and R. Péteri. A brief survey of dynamic texture description and recognition. *Proc. Conf. Computer Recognition Systems*, 5:17–26, 2005.

[7] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, , and T. Huang. Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In *IEEE Conf. Comp. Vision and Pattern Recognition*, volume 1, pages 595–601, 2003.

[8] J. Davis and A. Bobick. The representation and recognition of human movement using temporal templates. In *IEEE Conf. Comp. Vision and Pattern Recognition*, pages 928–934, 1997.

[9] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. *Lecture Notes in Computer Science*, 4738:488–500, 2007.

[10] P. Ekman, W. Friesen, and J. Hager. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. A Human Face, Salt Lake City, UT, 2002.

[11] P. Ekman and E. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 2005.

[12] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[13] S. Gokturk, J. Bouguet, C. Tomasi, and B. Girod. Model-based face tracking for viewindependent facial expression recognition. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 272–278, 2002.

[14] G. Guo and C. Dyer. Learning from examples in the small sample case - face expression recognition. *IEEE Trans. Systems, Man and Cybernetics*, 35(3):477–488, 2005.

[15] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 46–53, 2000.

[16] S. Koelstra and M. Pantic. Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 1–8, 2008.

[17] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. Image Processing*, 16(1):172–187, 2007.

[18] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[19] Z. Lu, W. Xie, J. Pei, and J. Huang. Dynamic texture recognition by spatio-temporal multiresolution histograms. In *IEEE Workshop on Motion and Video Computing*, volume 2, pages 241–246, 2005.

[20] S. Lucey, A. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 275–286. I-Tech Education and Publishing, Vienna, 2007.

[21] M. Pantic and M. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, 2007.

[22] M. Pantic and I. Patras. Detecting Facial Actions and their Temporal Segments in Nearly Frontal-View Face Image Sequences. In *Proc. IEEE Conf. Systems, Man and Cybernetics*, volume 4, pages 3358–3363, 2005.

[23] M. Pantic and I. Patras. Dynamics of facial expressions - recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Systems, Man and Cybernetics*, 36(2):433–449, 2006.

[24] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: a survey. *Lecture Notes on Artificial Intelligence*, 4451:47–71, 2007.

[25] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions - the state of the art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[26] M. Pantic and L. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Trans. Systems, Man and Cybernetics*, 34(3):1449–1461, 2004.

[27] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. IEEE Conf. Multimedia and Expo*, pages 317–321, 2005.

[28] R. Polana and R. Nelson. Temporal texture and activity recognition. *Motion-Based Recognition*, pages 87–115, 1997.

[29] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes. Nonrigid registration using free-form deformations: Application to breast mr images. *IEEE Transactions on medical imaging*, 18(8):712–721, 1999.

[30] P. Saisan, G. Doretto, Y. Wu, and S. Soatto. Dynamic texture recognition. In *IEEE Conf. Comp. Vision and Pattern Recognition*, volume 2, pages 58–63, 2001.

[31] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):1–19, 2001.

[32] Y. Tian, T. Kanade, and J. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 218–223, 2002.

[33] Y. Tong, W. Liao, and Q. Ji. Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[34] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. *IEEE Conf. Comp. Vision and Pattern Recognition*, 3:149, 2006.

[35] M. Valstar and M. Pantic. Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics. *Lecture Notes on Computer Science*, 4796:118–127, 2007.

[36] M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection from face video. In *Proc. IEEE Conf. Systems, Man and Cybernetics*, pages 635–640, 2004.

[37] D. Vukandinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Proc. IEEE Conf. Systems, Man and Cybernetics*, volume 2, pages 1692–1698, 2005.

[38] Z. Wen and T. Huang. Capturing subtle facial motions in 3d face tracking. In *Proc. Int'l. Conf. Computer Vision*, volume 2, page 1343, 2003.

[39] J. Whitehill and C. Omlin. Haar features for FACS AU recognition. In *Proc. IEEE Intl Conf. Face and Gesture Recognition*, pages 97–101, 2006.

[40] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. In *Proc. ACM Conf. Multimodal Interfaces*, pages 126–133, 2007.

[41] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[42] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequence. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.

[43] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[44] G. Zhao and M. Pietikäinen. Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. *Pattern recognition letters*, 30(12):1117–1127, September 2009.

Sander Koelstra (S' 2009) Sander Koelstra received the B.Sc. and M.Sc. degrees in Computer Science from the Delft University of Technology, The Netherlands, in 2006 and 2008, respectively. He is currently a PhD student with the Department of Electronic Engineering in the Queen Mary University of London. His research interests lie in the areas of computer vision, brain-computer interaction and pattern recognition.

Maja Pantic Maja Pantic received the MSc and PhD degrees in Computer Science from the Delft University of Technology, The Netherlands, in 1997 and 2001, respectively. She is currently a reader in Multimodal HCI in the Department of Computing, Imperial College London, and a professor of affective and behavioral computing in the Department of Computer Science, University of Twente. She is the Editor in Chief of Image and Vision Computing Journal and an associate editor for the IEEE Transactions on Systems, Man, and Cybernetics Part B. She is a guest editor, organizer, and committee member of more than 10 major journals and conferences. Her research interests include computer vision and machine learning applied to face and body gesture recognition, multimodal human behaviour analysis, and context-sensitive human-computer interaction (HCI).

Ioannis (Yiannis) Patras (S' 1997, M'2002) received the B.Sc. and M.Sc. degrees in computer science from the Computer Science Department, University of Crete, Heraklion, Greece, in 1994 and in 1997, respectively, and the Ph.D. degree from the Department of Electrical Engineering, Delft University of Technology, The Netherlands, in 2001. He has been a Postdoctorate Researcher in the area of multimedia analysis at the University of Amsterdam, and a Postdoctorate Researcher in the area of vision-based human machine interaction at TU Delft. Between 2005 and 2007 was a Lecturer in Computer Vision at the Department of Computer Science, University of York, York, UK. Since 2007 he is a Lecturer in Computer Vision in the Department of Electronic Engineering in the Queen Mary University of London. He is/has been in the organizing committee of IEEE SMC 2004 and of Face and Gesture Recognition 2008 and is the general chair of WIAMIS 2009. He is associate editor in the Image and Vision Computing Journal and in the Journal of Multimedia. His research interests lie in the areas of computer vision and pattern recognition, with emphasis on motion analysis, and their applications in multimedia data management, multimodal human computer interaction, and visual communications. Currently, he is interested in the analysis of Human Motion, including the detection, tracking and understanding of facial and body gestures.