Recognition of Genuine Smiles

Hamdi Dibeklioğlu, *Member, IEEE*, Albert Ali Salah, *Member, IEEE*, and Theo Gevers, *Member, IEEE*

***Abstract*—Automatic distinction between genuine (spontaneous) and posed expressions is important for visual analysis of social signals. In this paper, we describe an informative set of features for the analysis of face dynamics, and propose a completely automatic system to distinguish between genuine and posed enjoyment smiles. Our system incorporates facial landmarking and tracking, through which features are extracted to describe the dynamics of eyelid, cheek, and lip corner movements. By fusing features over different regions, as well as over different temporal phases of a smile, we obtain a very accurate smile classifier. We systematically investigate age and gender effects, and establish that age-specific classification significantly improves the results, even when the age is automatically estimated. We evaluate our system on the 400-subject UvA-NEMO database we have recently collected, as well as on three other smile databases from the literature. Through an extensive experimental evaluation, we show that our system improves the state of the art in smile classification and provides useful insights in smile psychophysics.**

***Index Terms*—Affective computing, expression dynamics, expression spontaneity, face analysis, genuine smile, human-computer interaction, social signals.**

I. Introduction

HUMAN facial expressions are indispensable elements of non-verbal communication. Since faces can reveal the mood or the emotional feeling of a person, automatic understanding and interpretation of facial expressions provide a natural way to interact with computers. Automatic analysis and classification of emotional facial expressions have been an active research topic since the Facial Action Coding system (FACS) was proposed by Ekman [1]. Since the literature on automated facial expression recognition is extensive, we refer the reader to two comprehensive surveys [2], [3].

In recent studies, analysis of spontaneous facial expressions have gained more interest. For social interaction analysis, it is necessary to distinguish genuine (spontaneous/felt) expressions from the posed (deliberate) ones since they convey different meanings. Spontaneous expressions can reveal states of attention, agreement and interest, as well as deceit. The foremost facial expression for spontaneity analysis is the smile as it is the most frequently performed expression. A smile can signal enjoyment, embarrassment, politeness, etc. [4]. It is also used to mask other emotional expressions, since it is the easiest emotional facial expression to pose voluntarily [5], [6].

Several characteristics of genuine and posed smiles, such as symmetry, speed, and timing are analyzed in the literature [7]–[9]. Their findings suggest that different facial regions contribute differently to the classification of smiles. In this paper, we combine region-specific movement dynamics (e.g. duration, amplitude, speed and acceleration) to detect the genuineness of enjoyment smiles. To this end, we propose a generic set of features that can be applied to different facial regions. We assess the discrimination power of regional dynamics, and demonstrate that the eye region contains the most useful information for distinguishing between spontaneous and posed smiles automatically.

Our contributions are: 1) we report the most extensive set of comparative results on automatic smile analysis, using the largest spontaneous/posed enjoyment smile database in the literature, as well as several older databases; 2) we report an accurate smile classification method, which outperforms the state-of-the-art methods; 3) we provide new empirical findings on age related differences in smile expression dynamics; 4) we provide region-specific analysis of facial feature movements under various conditions; and 5) we systematically explore different factors influencing smile classification, including the contributions of different facial regions and temporal phases, age and gender. Since we compare our proposed method with three recent approaches from the literature, on four different databases, our results accurately depict the state of the art in smile analysis.

A preliminary version of this paper appeared as [10]. Apart from increasing the level of detail throughout the paper, the contributions over our earlier work can be listed as: 1) an online analysis system is implemented and assessed; 2) weighted SUM fusion is included in the analysis, improving the results; 3) gender effects are analyzed; 4) age effects are analyzed in detail, additional experiments performed with automatic age estimation; 5) a number of experiments are conducted to assess the contribution of each feature, phase and region; 6) related work and databases are expanded, new experimental results are reported on the BBC, MMI, and SPOS databases; 7) facial fea-

Manuscript received February 04, 2014; revised June 10, 2014 and December 18, 2014; accepted December 25, 2014. Date of publication January 22, 2015; date of current version February 12, 2015. This work was supported by the Dutch national program COMMIT and by Boğaziçi University under Project BAP-6531. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

H. Dibeklioğlu is with Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Amsterdam 1098 XH, The Netherlands, and is also with the Pattern Recognition and Bioinformatics Group, Delft University of Technology, Delft 2628 CD, The Netherlands (e-mail: h.dibeklioglu@tudelft.nl).

A. A. Salah is with the Computer Engineering Department, Boğaziçi University, Istanbul 34342, Turkey (e-mail: salah@boun.edu.tr).

T. Gevers is with Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Amsterdam 1098 XH, The Netherlands, and is also with the Computer Vision Center, Universitat Autónoma de Barcelona, Barcelona 08193, Spain (e-mail: th.gevers@uva.nl).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes a video, which demonstrates how online analysis would fare in distinguishing between posed and spontaneous (genuine) enjoyment smiles. This material is 1.6 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2394777

ture tracking is separately assessed; and 8) different classifiers are assessed.

This paper is structured as follows. In Section II, related work in smile and spontaneity analysis is given. Section III describes the proposed method for smile classification, followed by Section IV that describes the UvA-NEMO Smile Database and other spontaneous/posed smile databases. Then, Section V presents our experimental results, separated into subsections that analyze each factor in the smile classification system separately. In Section VI, the findings of this study are discussed. Section VII concludes the paper.

II. Related Work

This section first summarizes the physiognomy of smiles, and then reports related work on automatic smile analysis.

A. The Physiognomy of Smiles

The smile is the easiest emotional facial expression to pose voluntarily [6]. Broadly, a smile can be identified as the upward movement of the lip corners, which corresponds to Action Unit 12 (AU12) in the facial action coding system (FACS) [1]. In terms of anatomy, the *zygomatic major* muscle contracts and raises the corners of the lips during a smile [8]. In terms of dynamics, smiles are composed of three non-overlapping phases; the onset (neutral to expressive), apex, and offset (expressive to neutral), respectively. Ekman individually identified 18 different smiles (such as enjoyment, fear, miserable, embarrassment, listener response smiles) by describing the specific visual differences on the face and indicating the accompanying action units, however temporal dynamics for each smile type were not described [6].

Empirical research into the physiognomy of smiles started with Guillaume Duchenne in the mid-nineteenth century. Duchenne experimented on muscle activities during smiles, and proposed that smiles resulting from felt joy not only utilize the *zygomaticus major* muscle, but also the *orbicularis oculi* (a circular muscle around the eyes). Duchenne claimed that the *orbicularis oculi* could not be controlled voluntarily during posed smiles [11]. This kind of joy smiles are called Duchenne smiles (D-smiles) in his honor. After more than a century, Ekman and Friesen supported Duchenne's observations for felt smiles of positive emotions, with empirical findings [8]. In [12], a strong correlation between D-smiles and felt enjoyment smiles were found. However, the definition of D-smiles was updated as the combined contraction of *zygomaticus major* and the outer strands (*pars lateralis*) of *orbicularis oculi*, since fewer people can voluntarily contract the outer strands of *orbicularis oculi*, as compared to its inner strands [13].

Contraction of the *orbicularis oculi, pars lateralis* raises the cheek, narrows the eye aperture, and forms wrinkles (crows-feet) on the external side of the eyes. This activation corresponds to Action Unit 6 (AU6) and is named as the Duchenne marker (D-marker) in the literature. [14] indicates that most people cannot voluntarily contract *orbicularis oculi, pars lateralis* and the ones who can do it usually cannot activate this muscle on both sides of their face simultaneously. However, new empirical findings question the reliability of the D-marker [15]–[17]. Recently, it has been shown that *orbicularis oculi, pars lateralis* can be active or inactive under both spontaneous and posed conditions with similar frequencies [18]. On the other hand, untrained people consistently use the D-marker to recognize genuine and posed enjoyment smiles [19].

B. Analysis of Smiles

Most of the previous studies regarded the D-marker as the most reliable evidence for detection of felt enjoyment smiles [8], [20]–[22]. But other cues are also considered in the literature. For instance, symmetry is potentially informative to distinguish genuine and posed enjoyment smiles [8]. In [7], it has been claimed that genuine enjoyment smiles are more symmetrical than posed ones. Later studies reported no significant difference [23], [15].

In the last decade, dynamical properties of smiles (such as duration, speed, and amplitude of smiles; movements of head and eyes) [9], [24], [25] received attention as opposed to morphological cues [26], [27] to discriminate between genuine and posed smiles. In [28], Cohn *et al.* analyze correlations between lip-corner displacements, head rotations, and eye motion during spontaneous smiles. In another study, Cohn and Schmidt report that spontaneous smiles have smaller onset amplitude of lip corner movement, but a more stable relation between amplitude and duration [9]. Other related findings show that the maximum speed of the smile onset is higher in posed samples [29]. Furthermore, the maximum speed of the smile onset is higher in posed samples and posed eyebrow raises have higher maximum speed and larger amplitude, but shorter duration than spontaneous ones [15].

From the perspective of computer vision, several temporal methods are proposed for automatic classification of genuine and posed facial expressions [9], [30], [24], [25], [31]. These studies not only use the defined differences of genuine/posed expressions, but also propose new cues for classification. In [9], Cohn and Schmidt propose a system which distinguishes spontaneous and deliberate enjoyment smiles by a linear discriminant classifier using duration, amplitude, and $\frac{\text{duration}}{\text{amplitude}}$ measures of smile onsets. They analyze the significance of the proposed features and show that the amplitude of the lip corner movement is a strong linear function of duration in spontaneous smile, but not in deliberate ones.

In [30], Valstar *et al.* propose a method to automatically discriminate between spontaneous and deliberate brow actions using intensity, duration, trajectory, symmetry, and occurrence order of the actions. In [24], a multimodal system is presented to classify posed and genuine smiles. GentleSVM-Sigmoid classifier is used with the fusion of shoulder, head and inner facial movements.

Recently, Dibeklioğlu *et al.* have proposed a system which uses eyelid movements to classify genuine and posed enjoyment smiles, where distance-based and angular features are defined in terms of changes in eye aperture [25]. Several classifiers are compared and the reliability of eyelid movements are shown to be superior to that of the eyebrows, cheek, and lip movements for smile classification.

In [31], Pfister *et al.* propose a spatio-temporal method using both natural and infrared face videos to discriminate between

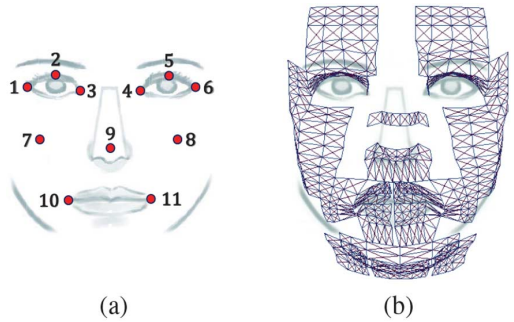

Fig. 1. (a) Used facial feature points with their indices and (b) the 3-D mesh model.

spontaneous and posed facial expressions. By enabling the temporal space and using the image sequence as a volume, they extend the Completed Local Binary Patterns (CLBP) texture descriptor into the spatio-temporal CLBP-TOP features for this task.

In conclusion, the most relevant facial cues for smile classification in the literature are: 1) the D-marker, 2) the symmetry, and 3) the dynamics of smiles. Instead of analyzing these facial cues separately, in this paper, the aim is to use a more generic descriptor set which can be applied to different facial regions to enhance the indicated facial cues with detailed dynamic features. Additionally, we focus on the dynamical characteristics of eyelid movements (such as duration, amplitude, speed, and acceleration), instead of simple displacement analysis, motivated by the findings of [9] and [25]. We report in Section V-I a comparison between the best performing methods in the literature.

III. Method

One of the contributions of this paper is an accurate smile classification approach. In this section, details of the proposed spontaneous/posed enjoyment smile classification system are summarized. The flow of the system is as follows. Facial fiducial points are located in the first frame, and tracked during the rest of the smile video. These points are used to calculate displacement signals of eyelids, cheeks, and lip corners. Onset, apex, and offset phases of the smile are estimated using the normalized displacement of the lip corners. Afterwards, descriptive features for eyelid, cheek, and lip corner movements are extracted from each phase. After a feature selection procedure, the most informative features with minimum dependency are used to train Support Vector Machine (SVM) classifiers.

A. Facial Feature Tracking

To analyze the facial dynamics, 11 facial feature points (eye corners, center of upper eyelids, cheek centers, nose tip, lip corners) are tracked in the videos [see Fig. 1(a)]. Note that, the cheek center is computed as the center location of the cheek patch [see Fig. 1(b)]. These fiducial points are specifically selected to describe the movements on the eye, cheek, and mouth regions, which are related to the most relevant facial cues in the literature for smile classification [13], [15]–[18]. Each point is initialized in the first frame of the videos for precise tracking and analysis. In our system, we use the piecewise Bézier volume deformation (PBVD) tracker, which is proposed by Tao and Huang [32] [see Fig. 1(b)]. We have introduced improved methods for its initialization, since it is fast and robust with accurate initialization [33].

The PBVD tracker employs a model-based approach. A 3-D mesh model of the face is constructed by warping the generic model to fit the facial features in the first frame of the image sequence. The generic face model consists of 16 surface patches. To form a continuous and smooth model, these patches are embedded in Bézier volumes. If $x(u,v,w)$ is a facial mesh point, then the Bézier volume [34] is defined as

$$x(u,v,w)=\sum_{i=0}^{n}\sum_{j=0}^{m}\sum_{k=0}^{l}b_{i,j,k}B_i^n(u)B_j^m(v)B_k^l(w) \qquad (1)$$

where points $b_{i,j,k}$ and variables $0<\{u,v,w\}<1$ control the shape of the volume. $B_i^n(u)$ denotes a Bernstein polynomial

$$B_i^n(u)=\binom{n}{i}u^i(1-u)^{n-i}. \qquad (2)$$

The expressions of $B_j^m(v)$ and $B_k^l(w)$ are similar. When the control points are moved, both the deformed volume and the displacement of x can be obtained using Equation (1). After fitting the face model, facial feature points (as well as head motion) can be tracked in 3-D according to the movement and the deformations of the mesh. To measure 2-D motion, template matching is used between frames at different resolutions. For more robust tracking, image templates of both the previous frame and the first frame of the sequence are used for matching. The estimated 2-D image motion is modeled as a projection of the 3-D movement onto the image plane. Then, the 3-D movement is calculated using projective motion of several points.

B. Feature Extraction

Three different face regions (eyes, cheeks, and mouth) are used to extract descriptive features. First of all, tracked 3-D coordinates of the facial feature points ℓ_i [see Fig. 1(a)] are used to align the faces in each frame. We estimate the 3-D pose of the face, and normalize the face with respect to roll, yaw, and pitch rotations. Since three non-colinear points are enough to construct a plane, we use three stable landmarks (eye centers and nose tip) to define a plane $\mathcal{P}$. Eye centers are defined as middle points between the inner and outer eye corners as $c_1=\frac{\ell_1+\ell_3}{2}$ and $c_2=\frac{\ell_4+\ell_6}{2}$. Angles between the positive normal vector $\mathcal{N}_\mathcal{P}$ of $\mathcal{P}$ and unit vectors U on X (horizontal), Y (vertical), and Z (perpendicular) axes give the relative head pose as follows:

$$\theta=\arccos\frac{U.\mathcal{N}_\mathcal{P}}{\|U\|\,\|\mathcal{N}_\mathcal{P}\|},\ \text{where}\ \mathcal{N}_\mathcal{P}=\overrightarrow{\ell_9c_2}\times\overrightarrow{\ell_9c_1}. \qquad (3)$$

In Equation (3), $\overrightarrow{\ell_9c_2}$ and $\overrightarrow{\ell_9c_1}$ denote the vectors from point ℓ_9 to points c_2 and c_1, respectively. $\|U\|$ and $\|\mathcal{N}_\mathcal{P}\|$ are the magnitudes of U and $\mathcal{N}_\mathcal{P}$ vectors. According to the face geometry, Equation (3) can estimate the exact roll (θ_z) and yaw (θ_y) angles of the face with respect to the camera. If we assume that the face is approximately frontal in the first frame, then the actual pitch angles (θ_x') can be calculated by subtracting the initial

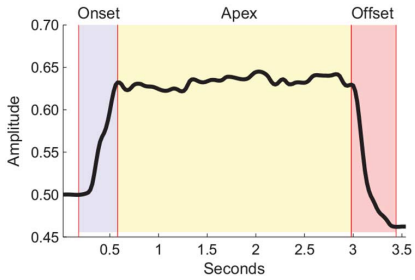

Fig. 2. Segmentation of temporal phases using the amplitude signal of the lip corners $\mathcal{D}_{\text{lip}}$.

value. Once the pose of the head is estimated, tracked points are normalized with respect to rotation, scale, and translation as follows:

$$\ell_i{}' = \left[\ell_i - \frac{c_1 + c_2}{2}\right] R \frac{100}{\rho(c_1, c_2)}, \quad R = R_x(-\theta_x')R_y(-\theta_y)R_z(-\theta_z) \tag{4}$$

where $\ell_i{}'$ is the aligned point and R_x, R_y, and R_z denote the 3-D rotation matrices for the given angles. $\rho()$ denotes the Euclidean distance. On the normalized face, the middle point between eye centers is located at the origin and the interocular distance (distance between eye centers) is set to 100 pixels. Since the normalized face is approximately frontal with respect to the camera, we ignore the depth (Z) values of the normalized feature points $\ell_i{}'$, and denote them as l_i.

After the normalization, the onset, apex, and offset phases of the smile are detected using the approach proposed in [35], by calculating the amplitude of the smile as the distance of the right lip corner to the lip center during the smile. Differently, we estimate the smile amplitude as the average amplitude of right and left lip corners, normalized by the length of the lip. Let $\mathcal{D}_{\text{lip}}(t)$ be the value of the normalized amplitude signal of the lip corners in the frame t. It is computed by

$$\mathcal{D}_{\text{lip}}(t) = \frac{\rho(\frac{l_{10}^1 + l_{11}^1}{2}, l_{10}^t) + \rho(\frac{l_{10}^1 + l_{11}^1}{2}, l_{11}^t)}{2\rho(l_{10}^1, l_{11}^1)} \tag{5}$$

where l_i^t denotes the 2-D location of the i^{th}point in frame t. The longest continuous increase in $\mathcal{D}_{\text{lip}}$ is defined as the onset phase. Similarly, the offset phase is detected as the longest continuous decrease in $\mathcal{D}_{\text{lip}}$. The phase between the last frame of the onset and the first frame of the offset defines the apex (see Fig. 2).

To extract features from the eyelids and the cheeks, additional amplitude signals are computed. We estimate the (normalized) eyelid aperture $\mathcal{D}_{\text{eyelid}}$ and cheek displacement $\mathcal{D}_{\text{cheek}}$ by

$$\mathcal{D}_{\text{eyelid}}(t) = \frac{\tau(\frac{l_1^t + l_3^t}{2}, l_2^t) + \tau(\frac{l_4^t + l_6^t}{2}, l_5^t)}{2\rho(l_1^t, l_3^t)}, \tag{6}$$

$$\mathcal{D}_{\text{cheek}}(t) = \frac{\rho(\frac{l_7^1 + l_8^1}{2}, l_7^t) + \rho(\frac{l_7^1 + l_8^1}{2}, l_8^t)}{2\rho(l_7^1, l_8^1)} \tag{7}$$

where $\tau(l_i, l_j) = \kappa(l_i, l_j)\rho(l_i, l_j)$, and $\kappa(l_i, l_j)$ denotes the relative vertical location function, which equals to -1 if l_j is located (vertically) below l_i on the face, and 1 otherwise. $\mathcal{D}_{\text{lip}}$, $\mathcal{D}_{\text{eyelid}}$, and $\mathcal{D}_{\text{cheek}}$ are hereafter referred to as amplitude signals. Extraction of the amplitude signals are visualized in Fig. 3. In addition to the amplitudes, speed $\mathcal{V}(t) = \frac{d\mathcal{D}}{dt}$, and acceleration $\mathcal{A} = \frac{d^2\mathcal{D}}{dt^2}$ signals are computed.

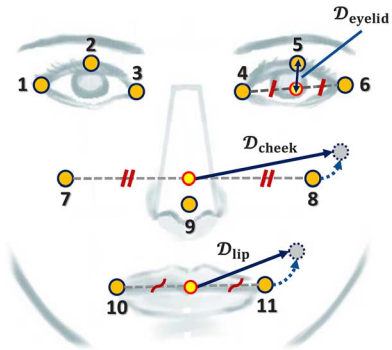

Fig. 3. Visualization of the amplitude signals, which are defined as the mean of left/right amplitude signals on the face. For simplicity, visualizations are shown on a single side of the face.

TABLE I
DEFINITIONS OF THE EXTRACTED FEATURES, AND THE RELATED FACIAL CUES WITH THOSE. THE RELATED FACIAL CUES ARE GIVEN BY SUPERINDICES, WHERE d, m, AND s DENOTE DYNAMICS, D-MARKER, AND SYMMETRY, RESPECTIVELY. THE RELATION WITH D-MARKER IS ONLY VALID FOR EYELID FEATURES

Feature	Definition
Duration^d:	$\left[\ \frac{\eta(\mathcal{D}^+)}{\omega}\ ,\ \frac{\eta(\mathcal{D}^-)}{\omega}\ ,\ \frac{\eta(\mathcal{D})}{\omega}\ \right]$
Duration Ratio^d:	$\left[\ \frac{\eta(\mathcal{D}^+)}{\eta(\mathcal{D})}\ ,\ \frac{\eta(\mathcal{D}^-)}{\eta(\mathcal{D})}\ \right]$
Maximum Amplitude^{d,m}:	$\max(\mathcal{D})$
Mean Amplitude^{d,m}:	$\left[\ \frac{\sum\mathcal{D}}{\eta(\mathcal{D})}\ ,\ \frac{\sum\mathcal{D}^+}{\eta(\mathcal{D}^+)}\ ,\ \frac{\sum\lvert\mathcal{D}^-\rvert}{\eta(\mathcal{D}^-)}\ \right]$
STD of Amplitude^d:	$\text{std}(\mathcal{D})$
Total Amplitude^d:	$\left[\ \sum\mathcal{D}^+\ ,\ \sum\lvert\mathcal{D}^-\rvert\ \right]$
Net Amplitude^d:	$\sum\mathcal{D}^+ - \sum\lvert\mathcal{D}^-\rvert$
Amplitude Ratio^d:	$\left[\ \frac{\sum\mathcal{D}^+}{\sum\mathcal{D}^+ + \sum\lvert\mathcal{D}^-\rvert}\ ,\ \frac{\sum\lvert\mathcal{D}^-\rvert}{\sum\mathcal{D}^+ + \sum\lvert\mathcal{D}^-\rvert}\ \right]$
Maximum Speed^d:	$\left[\ \max(\mathcal{V}^+)\ ,\ \max(\lvert\mathcal{V}^-\rvert)\ \right]$
Mean Speed^d:	$\left[\ \frac{\sum\mathcal{V}^+}{\eta(\mathcal{V}^+)}\ ,\ \frac{\sum\lvert\mathcal{V}^-\rvert}{\eta(\mathcal{V}^-)}\ \right]$
Maximum Acceleration^d:	$\left[\ \max(\mathcal{A}^+)\ ,\ \max(\lvert\mathcal{A}^-\rvert)\ \right]$
Mean Acceleration^d:	$\left[\ \frac{\sum\mathcal{A}^+}{\eta(\mathcal{A}^+)}\ ,\ \frac{\sum\lvert\mathcal{A}^-\rvert}{\eta(\mathcal{A}^-)}\ \right]$
Net Ampl., Duration Ratio^d:	$\frac{(\sum\mathcal{D}^+ - \sum\lvert\mathcal{D}^-\rvert)\omega}{\eta(\mathcal{D})}$
Left/Right Ampl. Difference^s:	$\frac{\lvert\sum\mathcal{D}_L - \sum\mathcal{D}_R\rvert}{\eta(\mathcal{D})}$

In summary, description of the used features and the related facial cues with those are given in Table I. Note that the defined features are extracted separately from each phase of the smile. As a result, we obtain three feature sets for each of the eye, mouth and cheek regions. Each phase is further divided into increasing ($^+$) and decreasing ($^-$) segments, for each feature set. This allows a more detailed analysis of the feature dynamics.

In Table I, signals symbolized with superindex ($^+$) and ($^-$) denote the segments of the related signal with continuous increase and continuous decrease, respectively. For example, $\mathcal{D}^+$ pools the increasing segments in $\mathcal{D}$ (see Fig. 4). η defines the length (number of frames) of a given signal, and ω is the frame

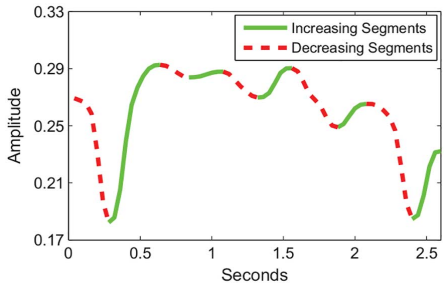

Fig. 4. Increasing and decreasing segments on an amplitude signal.

rate of the video. $\mathcal{D}_L$ and $\mathcal{D}_R$ define the amplitudes for the left and right sides of the face, respectively. For each face region, three 25-dimensional feature vectors are generated by concatenating these features.

In some cases, the features cannot be calculated. For example, if we extract features from the amplitude signal of the lip corners $\mathcal{D}_{\text{lip}}$ using the onset phase, then decreasing segments will be an empty set ($\eta(\mathcal{D}^-) = 0$). For such exceptions, all the features describing the related segments are set to zero.

C. Feature Selection and Classification

To classify genuine and posed smiles, individual SVM classifiers are trained for different face regions. As described in Section III-B, we extract three 25-dimensional feature vectors for each face region. To deal with feature redundancy, we use the Min-Redundancy Max-Relevance (mRMR) algorithm to select the discriminative features [36]. mRMR is an incremental method for minimizing the redundancy while selecting the most relevant information. Let S_{m-1} be the set of selected $m-1$ features, then the m^{th} feature can be selected from the set $\{F - S_{m-1}\}$ as

$$\max_{f_j \in F - S_{m-1}} \left[I(f_j, c) - \frac{1}{m-1} \sum_{f_i \in S_{m-1}} I(f_j, f_i) \right] \tag{8}$$

where I shows the mutual information function and c indicates the target class. F and S denote the entire feature set, and the selected features, respectively. Equation (8) is used to determine which feature is selected at each iteration of the algorithm. The size of the selected feature set is determined based on the validation error.

During the training of our system, both individual feature vectors for the onset, apex, offset phases, and a vector with their fusion are generated. The most discriminative features on each of the generated feature sets are selected using mRMR. Minimum classification error on a separate validation set is used to determine the most informative facial region, and the most discriminative features on the selected region. Similarly, to optimize the SVM configuration, linear, polynomial, and radial basis function (RBF) kernels with different parameters (size of RBF kernel, degree of polynomial kernel) are tested on the validation set. The test partition of the dataset is not used for parameter optimization.

[1][Online] Available: http://www.uva-nemo.org

IV. Database

A. UvA-NEMO Smile Database

We have recently collected the UvA-NEMO Smile Database[1][10] to analyze the dynamics of spontaneous/posed enjoyment smiles. Data collection was carried out as a part of Science Live, the innovative research programme of Science Center NEMO.[2] This database is composed of videos (in RGB color) recorded with a Panasonic HDC-HS700 3MOS camcorder, placed on a monitor, at approximately 1.5 meters away from the recorded subjects. Videos were recorded with a resolution of 1920 × 1080 pixels at a rate of 50 frames per second under artificial D65 daylight illumination. Additionally, a color chart is present on the background of the videos for color normalization. Fig. 5 shows sample frames from the UvA-NEMO Smile Database.

The database has 1240 smile videos (597 spontaneous, 643 posed) from 400 subjects (185 female, 215 male). The ages of subjects vary from 8 to 76 years, and there are 149 young people (235 spontaneous, 240 posed) and 251 adults (362 spontaneous, 403 posed). 43 subjects do not have spontaneous smiles and 32 subjects have no posed smile samples. (See Fig. 6 for age and gender distributions).

For posed smiles, each subject was asked to pose an enjoyment smile as realistically as possible, after being shown the proper way in a sample video. To elicit genuine enjoyment smiles, a set of short, funny video segments were shown to each subject for approximately five minutes. While subjects were watching the videos, their facial expressions were recorded. During this session, we did not interact or communicate with subjects. Duration of both posed and elicited (spontaneous) smiles were segmented by two trained annotators. Segments start/end with neutral or near-neutral expressions. To able to have a balanced number of spontaneous and posed smiles, maximum two posed and two spontaneous smiles were selected by seeking consensus of the two annotators. If more than two spontaneous/posed smiles were selected by the consensus of annotators, smiles that start with more frontal pose were included in the database.

The mean duration of the spontaneous and posed smile segments are 4.9 ($\sigma = 2.1$) seconds, and 3.1 ($\sigma = 0.9$) seconds, respectively. Average interocular distance on the database is approximately 200 pixels (estimated by using the tracked landmarks). 50 subjects wear eyeglasses.

B. Other Smile Databases

Facial expression databases in the literature rarely contain spontaneous smiles. We have used several existing databases (BBC, MMI, SPOS) to report results with the proposed method. This section will give the details of publicly available databases which have both spontaneous and posed smile/laughter content. Table II shows a comparative overview of these databases.

BBC Smile Dataset[3] was gathered from "Spot the fake smile" test on the BBC website. The dataset has 10 spontaneous and

[2][Online] Available: http://www.e-nemo.nl.

[3][Online] Available: http://www.bbc.co.uk/science/humanbody/mind/surveys/smiles/

Fig. 5. Sample frames from the UvA-NEMO Smile Database showing neutral face (top), posed enjoyment smile (middle), and spontaneous enjoyment smile (bottom).

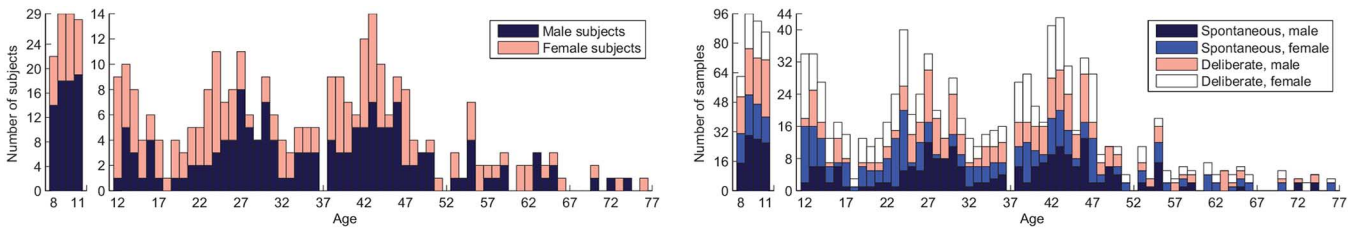

Fig. 6. Age and gender distributions for the subjects (left) and for the smiles (right) in the UvA-NEMO Smile Database.

TABLE II
DETAILS OF SMILE/LAUGHTER CONTENT IN DIFFERENT DATABASES. NOTE THAT THE POSED PART OF USTC-NVIE INCLUDES ONLY THE NEUTRAL AND THE MOST EXPRESSIVE (APEX) IMAGES

Database	Number of Subjects			Video		Age	
	Spon.	Posed	Total	Resolution	FPS	Annotation	Range
BBC	10	10	20	314 × 286	25	No	Unknown
MAHNOB	22	20	22	720 × 576	25	Some	Unknown
MMI	25	30	55	720 × 576 640 × 480	25 29	Yes	19–64
SPOS	7	7	7	640 × 480	25	No	Unknown
USTC-NVIE	133	104	148	640 × 480 704 × 480	30 30	No	17–31
UvA-NEMO	357	368	400	1920 × 1080	50	Yes	8–76

10 posed smile videos, each from a different subject and each starting and ending with a neutral face.

MAHNOB Laughter Database [37] contains audio, video and thermal video recordings of 22 subjects while watching funny videos. There are 563 spontaneous laughter episodes, 849 speech utterances, 51 posed laughs, 67 speech-laughter episodes and 167 other vocalizations annotated in the database.

MMI Facial Expression Database [38] is not specifically gathered for smile classification, but the annotated frontal recordings include 74 posed smiles from 30 subjects, as well as spontaneous smiles/laughters from 25 subjects. Spontaneous content of the database has two subsets. The first subset includes 383 manually annotated/cut segments (from 16 subjects) showing several affective displays. The second subset has 9 uncut recordings (including audio) containing 164 annotated laughters from 9 subjects. For our experiments, 120 spontaneous smiles that start and end with a neutral face (of 15 subjects) are selected from the first subset.

SPOS Corpus [31] consists of natural color and infrared videos of six basic facial expressions. Each expression for each subject has spontaneous and posed recordings. The database contains only the onset phases of expressions. There are 66 spontaneous and 14 posed smiles from seven subjects.

USTC-NVIE Database [39] consists of images and videos of six basic facial expressions and neutral faces. Images are recorded in both natural color and infrared, simultaneously, under three different illumination conditions. NVIE database has two parts: spontaneous part includes image sequences of onset phases, where the posed part consists of only the neutral and the most expressive (apex) images. The database has 302 spontaneous smiles (image sequences of onset phases) from 133 subjects, and 624 posed smiles (single apex images) from 104 subjects.

V. EXPERIMENTAL RESULTS

For a detailed evaluation of the proposed method, the UvA-NEMO smile database is used in our experiments. Furthermore, we provide results on the BBC, MMI and SPOS databases in Section V-C to V-I. In the proposed method, SVM classifiers are used to distinguish between genuine and posed enjoyment smiles, but to validate the reliability of SVM, comparison of different classifiers is also included in Section V-A.

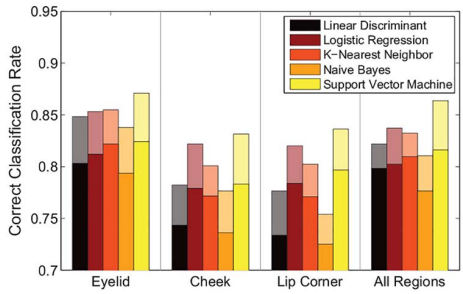

Fig. 7. Effect of using different models for classification with and without feature selection on the UvA-NEMO Smile Database. Upper (light-colored) part of the bars show the accuracy increase by enabling feature selection where lower parts denote the performance without feature selection.

We use a two level 10-fold cross-validation scheme: each time a test fold is separated, a 9-fold cross-validation is used to train the system, and parameters are optimized without using the test partition. There is no subject overlap between folds. A two level 7-fold cross-validation (leave one subject out) is used for the SPOS database, since it has only seven subjects.

In our experiments, we have tested linear, polynomial and RBF kernel SVMs with several parameters. The parameter set with the minimum validation error is selected and used for the related fold. In majority of the cases, the linear SVM was observed to result in the minimum validation error. For instance, during the training on all features of all regions, the linear SVM is chosen for seven of 10 folds. For a detailed analysis of the features and the system, tracking is initialized by manually annotated facial landmarks. The results with automatic initialization are also given in Sections V-H and V-I.

A. Assessment of Model and Feature Selection

To assess the robustness of different classification models, as well as the reliability of feature selection, we compare Linear Discriminant, Logistic Regression, k-Nearest Neighbor (k-NN), Naïve Bayes, and SVM classifiers trained both on selected and the whole set of features. The ideal number of nearest neighbors for k-NN is selected empirically by observing the validation error. To train the classifiers, we use the concatenated features of the onset, apex, and offset phases for each region, in addition to testing each phase individually. The UvA-NEMO Smile Database is used in this experiment.

As shown in Fig. 7, SVM outperforms the other methods on all regions, with and without feature selection. The accuracy of logistic regression and k-nearest neighbor follow that of SVM. If the features of all regions on each phase are concatenated, this combined feature does not improve the accuracy of classification over using only eyelid features (shown as 'All Regions' in Fig. 7). Even under these conditions, SVM still performs better than other classifiers.

When we analyze the results in terms of enabling feature selection, we observe that feature selection provides a 4.8% (relative) accuracy increase (on average), for different classifiers on individual regions. This improvement can be explained by discarding features which cause confusion. The highest relative improvement is obtained by SVM with an accuracy increase of 5.3%. Since these results confirm the usefulness of feature selection and SVM, they are used in the remainder of this section.

TABLE III
CORRECT CLASSIFICATION RATES ON THE UVA-NEMO SMILE DATABASE FOR DIFFERENT FACIAL REGIONS

Features	Classification Accuracy (%)			
	Onset	Apex	Offset	All Phases
Eyelid	78.79	68.39	63.47	87.10
Cheek	78.06	66.61	60.89	83.15
Lip Corner	82.58	65.65	54.11	83.63
All Regions	84.52	70.73	66.18	86.37

B. Assessment of Facial Regions and Temporal Phases

To evaluate the discriminative power of the eye, cheek, and mouth regions and temporal phases for smile classification, we use the features of onset, apex, and offset phases of all regions, individually. Features of all phases are also concatenated and tested for each region (shown as 'All Phases' in Table III). Additionally, concatenation of features for all regions is evaluated (shown as 'All Regions' in Table III). The UvA-NEMO Smile Database is used in these experiments.

For temporal phase segmentation, we employ a modified version of the approach proposed in [35], by using the change in lip corner displacements (see Section III-B). Different methods for temporal phase segmentation are proposed in a few other studies [40]–[43]. In [40], rule-based reasoning is used to segment temporal phases of different AUs. Similar to our approach, increase and decrease in lip corner displacements are used for AU12 segmentation. [41] employs multiclass SVMs and hybrid SVM-HMMs using distances between different landmark pairs to detect AU phases. However, costly manual annotation of temporal segments is required for training. [42] and [43] propose to model change in facial appearance for temporal phase segmentation. Yet, enabling appearance-based models introduces additional computational costs. In contrast to these approaches, we only use dynamics of point displacements.

As shown in Table III, the highest accuracy for individual phases is achieved by the onset features of lip corners (82.58%). However, the discriminative power of the apex and offset phases of lip corners do not reach those of the eyelids and cheeks. The most reliable apex and offset features are obtained from the eyelids, which provide the highest correct classification rate (87.10%) when the features of all phases are concatenated. Lip corners (83.63%) and cheeks (83.15%) follow the eyelids. The combined eyelid features have the minimum validation error in this experiment. When we analyze the results in terms of temporal phases, it is concluded that the onset is the most informative phase for all regions. It is followed by the apex and offset phases, respectively. The concatenation of all regional features for all phases does not improve the accuracy beyond using the combined eyelid features. This result shows that low-level fusion of features is not effective for smile classification, and mid-level or late fusion may be more promising. We assess these next.

TABLE IV
EFFECT OF DIFFERENT FUSION STRATEGIES ON THE CORRECT CLASSIFICATION RATES. NOTE THAT ONLY THE ONSET FEATURES ARE USED IN MID-LEVEL FUSION FOR THE SPOS DATABASE, SINCE IT DOES NOT INCLUDE APEX AND OFFSET PHASES

Fusion Strategy	Classification Accuracy (%)			
	BBC	MMI	SPOS	UvA-NEMO
Early	85.00	85.05	73.75	86.37
Mid-level, SUM	90.00	87.11	76.25	88.23
Mid-level, Weighted SUM	90.00	88.14	77.50	89.84
Mid-level, PRODUCT	90.00	86.60	75.00	87.66
Mid-level, Voting	90.00	85.05	76.25	88.87
Late, SUM	80.00	81.96	-	84.11
Late, Weighted SUM	85.00	83.51	-	85.48
Late, PRODUCT	85.00	85.05	-	85.81
Late, Voting	80.00	84.02	-	85.24

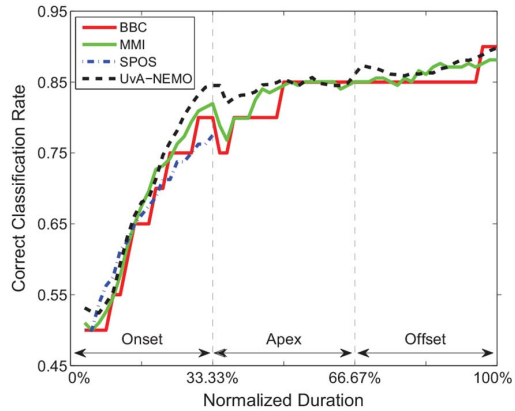

Fig. 8. Accuracy of the online system. To show the accuracy change with the usage of different amounts of smile duration, each phase is individually scaled to the same length (one third of the smile) in the visualization. Note that only the onset phase is used for SPOS database, since it does not include apex and offset phases.

C. Assessment of Fusion Strategies

To assess the performance of the different fusion techniques, three different strategies (early, mid-level, and late fusion) are defined and evaluated. Each fusion strategy enables feature selection before classification. In early fusion, features of onset, apex, and offset of all regions are fused into one low-abstraction vector and classified by a single classifier. Mid-level fusion concatenates features of all phases for each region, separately. Constructed feature vectors are individually classified by SVMs and the classifier outputs are fused. In the late fusion scheme, feature sets of onset, apex, and offset for all facial regions are individually classified by SVMs and regional classifier outputs are fused.

For the mid-level and late fusion strategies, classifier fusion is employed using one of SUM, weighted SUM, PRODUCT rule, or voting [44]. The SUM, weighted SUM, and PRODUCT rules fuse the computed posterior probabilities for the target classes. To estimate these posterior probabilities, sigmoids of SVM output distances are used. Weights for the weighted SUM rule are determined by the validation performance of the related classifiers. For voting, binary classification outputs of each classifier are counted as single votes, and the majority of votes determines the selected class.

As shown in Table IV, mid-level fusion provides the best performance across all databases, followed by early and late fusion, respectively. Elimination of redundant information by feature selection after low-level feature abstraction on each region, separately, and following higher level of abstraction for classification on different facial regions may explain the high accuracy of mid-level fusion.

D. Online Analysis of Temporal Information

A smile video from onset to offset contains a lot of frames. The system we have proposed gives a decision when the smile is completed, i.e. at the end of the offset phase. However, it may be necessary to give a decision while the smile is in progression. To understand how partial information would fare, we implement an online version of the proposed method. Mid-level fusion with weighted SUM rule is used as the fusion approach. The main difference of the online system is the temporal phase segmentation, which ordinarily detects onset, apex, and the offset phases. Given a smile segment, which starts from the onset (but may be incomplete), the duration of the lip corner's amplitude signal is analyzed. The longest continuous increase is selected as the onset phase. Afterwards, the system verifies whether the signal decreases in any part of the given signal. If there is a decreasing segment, then the longest continuous decrease is selected as the offset. The duration between the onset and offset is used as apex. In case of a stable duration after onset, it is selected as the apex without using the offset information.

Since the order of the temporal phases during a facial expression is fixed, the online system starts classification in the onset mode. When the apex is reached, it uses both onset and the apex in classification. In the final stage, all three phases are used. For these three modes, separate classifiers are trained. Since the proposed dynamics use speed and acceleration information, at least three frames are required in the latest phase to activate the classifier.

The performance of the online system is given in Fig. 8. To show the influence of using different portions of an evolving smile, each phase is individually scaled to the same length (one third of the smile) in the visualization. The results show that the correct classification accuracy decreases 5.26%, 5.09%, and 2.5% (absolute) in the beginning of the apex phase for BBC, MMI, and UvA-NEMO databases, respectively. The reason behind this is the incorrect detections of the apex phase. Online detection of the apex during the initial frames of the apex is a challenging problem, since the decrease in the amplitude signal after the onset phase may be easily confused with the offset.

Results for the UvA-NEMO Smile Database also show that even the half of the onset duration provides a correct classification rate of approximately 68%, where the accuracy reaches 84.52% and 86.13% at the end of onset and apex phases, respectively. Correct classification rate is 89.84% with the use of the entire smile duration. The onset is known to be the most informative phase for smile classification. However, the apex and offset increase the performance by 1.9% and 4.3% over the use of the onset and onset + apex, respectively. Results for other

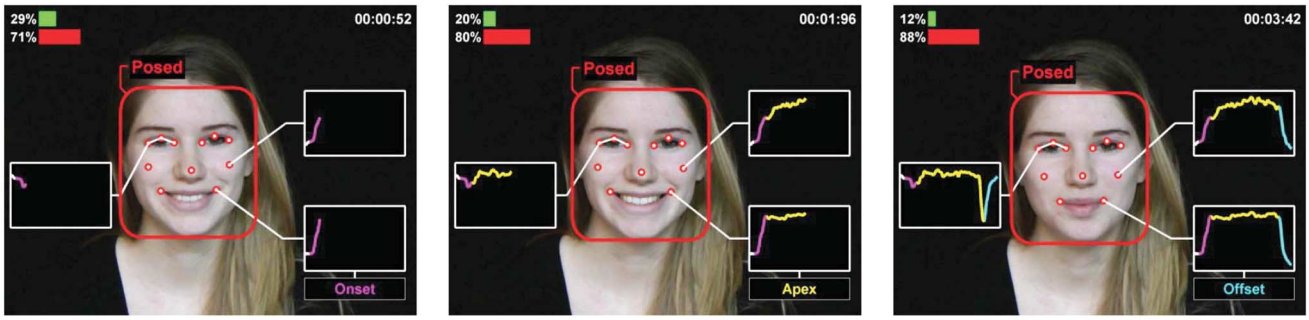

Fig. 9. Sample outputs of the online system during (a) onset, (b) apex, and (c) offset phases. Red (lower) and green (upper) bars with percentages show the probability of being posed and spontaneous, respectively. Signal plots show the regional amplitudes, and the temporal segmentation (best viewed in color).

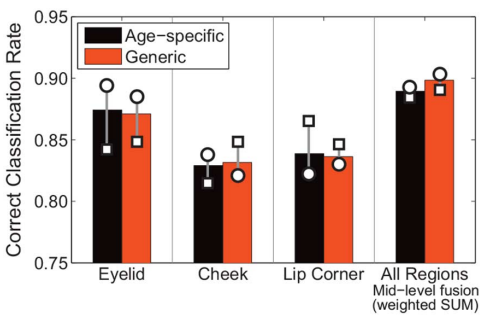

Fig. 10. Comparison of age-specific and generic methods for different regions on the UvA-NEMO Smile Database. Bars show the mean accuracy of the related method, where square and circle markers indicate the classification rates for *young people* and *adults*, respectively.

databases are also in line with these findings. Sample outputs of the online system during the onset, apex, and offset frames are shown in Fig. 9.

E. Effect of Age

The features on which we base our analysis may depend on the age of the subjects. To this end, we analyze the effect of age through two different experiments. In the first experiment, we split the UvA-NEMO Smile Database into two partitions as *young people* ($age < 18$ years), and *adults* ($age \geq 18$ years). All training and evaluation is repeated separately for the two partitions of the database. Since rapid craniofacial development during childhood and adolescence slows down, and facial structure is stabilized around the age of 18 years [45], the boundary age for data partitioning (*young people* and *adults*) is set as 18. The true ages (ground-truth) of subjects are used for this separation, and the resulting approach is hereafter referred to as the age-specific method. In the second experiment, we analyze the influence of including age information in the feature set using different databases. For this purpose, age information is added into each regional feature vector and the new feature set is evaluated using mid-level fusion with weighted SUM. Age information is defined and tested in two different ways: 1) age of subjects are used as they are; 2) age of subjects are grouped into bins of 10 years (1–10, 11–20, . . ., 71–80), and the resulting group labels (1, 2, . . ., 8) are included in the feature set.

Fig. 10 shows the classification accuracy of age-specific and generic methods for different facial regions on the UvA-NEMO Smile Database. Regional performances are given using the fused (onset, apex, offset) features of the related region. Mid-level fusion (weighted SUM) accuracies are also given in Fig. 10. Results show that the age-specific method performs better than the generic one when the eyelid or lip corner features are used. However, when all regions are used, the generic method reaches an accuracy of 89.84%, where the age-specific method reaches 88.95%. This difference is not statistically significant. When we analyze the regional results, it can be derived that the eyelid features perform better on *adults* (for both generic and age-specific methods) than cheeks and lip corners. On *young people*, the age-specific and generic methods with the eyelid features are approximately 5% and 4% less accurate compared to *adults*, respectively. On the other hand, the lip corner features using the age-specific method (86.53%), and cheek features using the generic method (84.84%) provide the highest regional accuracies for *young people*, respectively.

TABLE V
EFFECT OF USING AGE INFORMATION ON CLASSIFICATION ACCURACY

Method	Features	Classification Accuracy (%)			
		BBC	MMI	SPOS	UvA-NEMO
Age-specific	Only Dynamics	-	-	-	88.95
Generic	Only Dynamics	90.00	88.14	77.50	89.84
Generic	Dynamics + True Age	-	89.69	-	92.02
Generic	Dynamics + True Age Group	-	90.21	-	92.90
Generic	Dynamics + Est. Age	90.00	88.14	77.50	91.13
Generic	Dynamics + Est. Age Group	90.00	89.69	78.75	92.10

The results of the second experiment are given in Table V. Our results show that the use of true age (ground-truth) increases the accuracy of the proposed (generic) system by 1.55% and 2.19% (absolute), where using group labels of real ages provides an increase of 2.18% and 3.06% (absolute) for MMI and UvA-NEMO databases, respectively. Age ground truth is missing from BBC and SPOS, and the age specific system can only be implemented for UvA-NEMO, as all the other sets have only adults (visually confirmed). These account for the missing cells in Table V.

We also test the performance of the system using automatically estimated ages and age group labels (see the last two rows of Table V). Although quite a few number of automatic age estimation methods exist, we employ the method proposed by Dibeklioğlu *et al.* [46] to take advantage of using videos (instead

of static images) by enabling facial dynamics in the analysis. [46] combines the facial appearance and expression dynamics, and performs with a mean absolute error of 4.81($\pm$4.87) years on the UvA-NEMO smile database. This method uses the proposed features in this paper to describe facial expression dynamics and fuses them with Local Binary Patterns (LBP) descriptors, extracted from the first frame of the onset. Since the literature on automatic age estimation is extensive, we refer the reader to [47] and to the more recent [48] for related approaches.

Depending on the estimation error, automatic estimation of age does not improve the classification accuracy as much as using the true age. However, the use of the estimated age group labels in the feature set provides an accuracy increase of 1.27% (absolute) on average. These results clearly show that enabling age information in spontaneous/posed smile classification noticeably increases the accuracy. We test this statistically (using t-test analysis), and verify that the use of age groups instead of using the exact ages performs better, and significantly ($p < 0.05$) improves the accuracy over the sole use of dynamics.

The relation between aging and facial expressions is rarely analyzed in the literature, although several approaches for age estimation and facial expression recognition rely on similar features and classification paradigms. Recently, Guo *et al.* have analyzed the appearance differences of facial expressions for different age groups [49]. Their findings show that elderly people display facial expressions in a more subtle way in comparison to young people. Moreover, it is reported that aging can make differences in facial expression appearance based on the wrinkles and reduction in facial muscle elasticity. Due to these differences, they propose to detect facial expressions of different age groups as independent classes, and suggest reducing facial aging effects before the expression analysis. In [50] and [51], significant effects of expressions on age estimation accuracy have been shown in a quantitative manner. However, all these methods focus on facial appearance. Our previous studies, on the other hand, tackle the relations between temporal expression dynamics and aging, in automatic analysis of both age and facial expressions [10], [46].

F. Effect of Gender

To evaluate the effect of gender on the proposed features, we conduct two experiments similar to those reported in Section V-E. In the first experiment, we implement gender-specific and generic methods. For the gender-specific method, the UvA-NEMO Smile Database is split into two partitions as *males* and *females* (using gender ground-truth). All training and evaluation is repeated separately for the two partitions of the database. In the second experiment, gender labels are added into each regional feature vector and the new feature set is evaluated using mid-level fusion with weighted SUM.

Fig. 11 shows the comparison of gender-specific and generic methods on different regions. Both methods perform better on *males* in comparison to *females*. Best regional performances on both *males* and *females* are achieved by eyelid features. The eyelid and lip corner features provide higher correct classification rates for the generic method, but the gender-specific method performs slightly better on cheek features. When all regions are used, the generic method significantly ($p < 0.05$, using t-test analysis) outperforms the gender-specific one with an absolute accuracy increase of 2.58%.

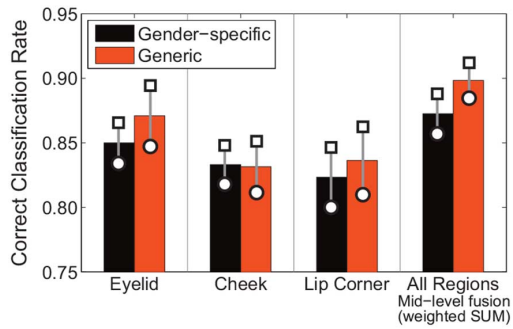

Fig. 11. Comparison of gender-specific and generic methods for different regions on the UvA-NEMO Smile Database. Bars show the mean accuracy of the related method, where square and circle markers indicate the classification rates for *male* and *female* subjects, respectively.

TABLE VI
EFFECT OF USING GENDER INFORMATION ON CLASSIFICATION ACCURACY

Method	Features	Classification Accuracy (%)			
		BBC	MMI	SPOS	UvA-NEMO
Gender-specific	Only Dynamics	85.00	86.08	73.75	88.95
Generic	Only Dynamics	90.00	88.14	77.50	89.84
Generic	Dynamics + True Gender	90.00	87.63	77.50	92.02

TABLE VII
INCLUDED FEATURES IN DIFFERENT GROUPS OF DYNAMICS. ($^-$,$^+$) DENOTE THAT THE RELATED FEATURE IS EXTRACTED FROM DECREASING OR INCREASING SEGMENTS, INDIVIDUALLY

Amplitude	**Amplitude/Duration**	**Duration**
Ampl. Ratio ($^-$,$^+$)	Mean Ampl.	Dur. Ratio ($^-$,$^+$)
Max. Ampl.	Mean Ampl. ($^-$,$^+$)	Dur. ($^-$,$^+$)
STD Ampl.	Net Ampl., Dur. Ratio	Total Dur.
Total Ampl. ($^-$,$^+$)	Symmetry	
Net Ampl.		
Speed	**Acceleration**	
Max. Speed ($^-$,$^+$)	Max. Accel. ($^-$,$^+$)	
Mean Speed ($^-$,$^+$)	Mean Accel. ($^-$,$^+$)	

In the second experiment, we analyze the effect of using gender labels in the feature set. As given in Table VI, gender-specific methods do not have a significant effect on the performance, but decrease the classification accuracy by 2.92% (absolute) on average in comparison to the generic method. On the other hand, using the gender labels in the feature set increases the accuracy by 3.34% (absolute) on average compared to the gender-specific system.

There are few computational studies that assess the informativeness of gender information in facial expression analysis [52], [27]. In [52], feature vectors obtained by Active Appearance Models are given to SVM cascades for estimating facial expressions and gender. Then, using the estimated gender labels, it is shown that the gender-specific expression recognition performs better than a generic approach. [27] shows that male subjects have more discriminative geometric features (distances

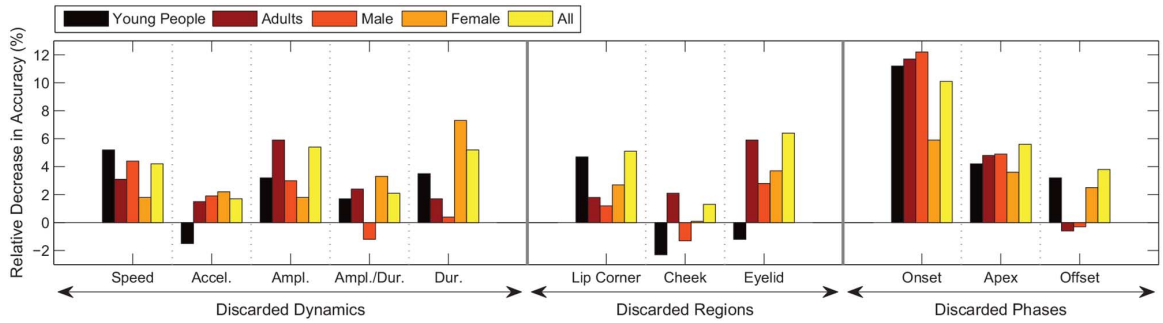

Fig. 12. Decrease in accuracy by discarding different feature groups. Negative values show relative increase in accuracy.

between different landmark pairs) than female subjects for distinguishing between spontaneous and posed expressions. Both approaches use static images and do not consider gender-specific differences in temporal dynamics.

G. Feature Analysis

In this section we analyze the informativeness of the proposed features, in order to provide more insight on the patterns of smile dynamics. First of all, we systematically eliminate different feature sets from the analysis, and observe the effects. To this end, we have grouped features by facial regions (lip corner, cheek, eyelid), temporal phases (onset, apex, offset), and the type of dynamics (speed, acceleration, amplitude, amplitude/duration, and duration). The included features in different groups of dynamics are given in Table VII. To obtain feature sets with similar size, we include symmetry and mean amplitude features in the Amplitude/Duration group. This is done based on the use of duration in the computation of symmetry and mean amplitude values.

We run our system by leaving out each of these groups, one at a time, and observe the effects on classification accuracy. Feature selection is disabled in this experiment for a direct comparison. Mid-level fusion with weighted SUM rule is used as the fusion approach. For each condition, the relative decrease in accuracy (with respect to using all features) is computed. Analysis is repeated for age and gender subsets, as well as for the whole database. For these experiments, we have combined BBC, MMI, SPOS, and UvA-NEMO databases into one set. Performing the same analysis on UvA-NEMO alone gives very similar results.

As shown in Fig. 12, the highest accuracy decrease is observed when the onset features are discarded from the analysis. In the temporal phase category, apex follows onset in importance, and offset is the least informative phase. In terms of regions, the eyelids give the most informative features on the whole database (shown as *All*), followed by lip corner and cheek features, respectively. However, the lip region is the most important one for young people, and discarding eyelid features may even improve the accuracy. For adults, cheek features are more effective than lip corners. When we evaluate dynamics over the whole database, amplitude, duration and speed features are the most effective features, respectively. On the other hand, the duration group is the most important one for females. For males and young people, discarding speed features causes the largest accuracy decrease. It should not be surprising that age and gender play a role in facial feature analysis. It is known that these factors play a major role in the morphology and appearance of the face.

As a next step, we use analysis of variance (ANOVA) to find out the individual differences between spontaneous and posed smiles, as well as the differences between subject groups. The results show that the most significant differences ($p < 0.001$, $\eta^2 > 0.12$) between young people and adults are in the maximum speed and the maximum acceleration of both eye closure and lip corner movements during the onset phase of smiles. When the most significant ($p < 0.001$, $\eta^2 > 0.035$) feature differences between males and females are analyzed, it is seen that maximum and mean apertures of eyes are significantly larger for females during onset, apex and offset phases. Such differences can explain the deviation of the feature effects for different gender and age subsets (also see Fig. 12 for classification).

10 highly significant feature differences ($p < 0.001$, $\eta^2 > 0.10$) between spontaneous and posed smiles with the largest effect sizes (η^2) on the whole set, and the change in effect size for these features on age and gender subsets are given in Table VIII. Note that these features are among the most frequently selected features by the mRMR algorithm. Our findings on the whole database show that maximum/mean speed and acceleration of eye closure and lip corner movements during smile onset are higher for posed smiles. Both the increasing amplitude duration, and the total duration of lip corner movements in smile apex are longer for spontaneous smiles.

When we analyze the deviation of effect sizes for different subsets, it is seen that eyelid features are less informative for young people in comparison to adults. Since the speed and acceleration of the eyelid movements are higher for posed smiles, faster eyelid movements of young people can cause confusion with posed smiles. On the other hand, slower lip corner movements of adults can cause the decrease in the effect size based on lower speed and acceleration of the lip movements during spontaneous smiles in comparison to posed ones. Between gender groups, deviations in effect size of different features are much less than those between age groups.

H. Effect of Tracking Initialization

In this paper, we use facial expression dynamics for smile classification, which are extracted using the displacement of facial landmarks. To assess the effect of tracking initialization on the accuracy, we use both manually and automatically annotated facial landmarks to initialize facial tracking. For the automatic

TABLE VIII
FEATURES WITH THE HIGHEST EFFECT SIZE ON THE WHOLE DATABASE AND THE CHANGE IN η^2 FOR DIFFERENT SUBSETS

Order	Features			Effect Size (η^2)				
	Region	Phase	Type	Young People	Adults	Males	Females	All
1	Lip Corner	Onset	Max Speed (Increasing Segments)	0.2094	0.1681	0.1966	0.1654	0.1812
2	Eyelid	Onset	Max Speed (Decreasing Segments)	0.1284	0.1987	0.1713	0.1833	0.1801
3	Lip Corner	Onset	Mean Speed (Increasing Segments)	0.2135	0.1619	0.1798	0.1760	0.1789
4	Eyelid	Onset	Mean Speed (Decreasing Segments)	0.1468	0.1759	0.1620	0.1798	0.1663
5	Eyelid	Onset	Max Accel. (Decreasing Segments)	0.1450	0.1654	0.1681	0.1543	0.1592
6	Lip Corner	Onset	Max Accel. (Increasing Segments)	0.1708	0.1491	0.1630	0.1528	0.1559
7	Lip Corner	Onset	Mean Accel. (Increasing Segments)	0.1524	0.1295	0.1496	0.1285	0.1364
8	Lip Corner	Apex	Duration (Increasing Segments)	0.1245	0.1429	0.1452	0.1316	0.1324
9	Eyelid	Onset	Mean Accel. (Decreasing Segments)	0.1294	0.1394	0.1415	0.1311	0.1302
10	Lip Corner	Apex	Total Duration	0.1215	0.1312	0.1440	0.1189	0.1297

TABLE IX
EFFECT OF MANUAL AND AUTOMATIC INITIALIZATION FOR TRACKING ON THE UVA-NEMO SMILE DATABASE

Features	Classification Accuracy (%)	
	Manual Init.	Auto Init.
Eyelid	87.10	85.73
Cheek	83.15	81.25
Lip Corner	83.63	82.66
All Regions	89.84	87.82

TABLE X
EFFECT OF MANUAL AND AUTOMATIC INITIALIZATION FOR TRACKING ON DIFFERENT DATABASES

Features	Classification Accuracy (%)	
	Manual Init.	Auto Init.
BBC	90.00	90.00
MMI	88.14	86.08
SPOS	77.50	75.00
UvA-NEMO	89.84	87.82

detection of landmarks, we use the state-of-art facial landmark detection system proposed by Dibeklioğlu [33]. This method models Gabor wavelet features of a neighborhood of landmarks using incremental mixtures of factor analyzers and enables a shape prior to ensure the integrity of the landmark constellation. It follows a coarse-to-fine strategy. Landmarks are initially detected on a coarse level and then fine-tuned for higher resolution. The mean localization error for the related landmarks [eye corners, center of upper eyelids, nose tip, lip corners, see Fig. 1(a)] is 3.96% (± 3.14) of the inter-ocular distance to the actual location of the landmarks. Actual locations of these landmarks for the first frames of each smile video in the UvA-NEMO Smile Database have been manually annotated by us. XM2VTS [53] and Bosphorus [54] databases have been used to train the landmarker, while AR database [55] has been employed for validation. Linear correlation coefficients between the extracted amplitude signals with manual and automatic initializations range between 0.93 and 1.00. To this end, the coefficients have been computed for each smile in UvA-NEMO database.

As shown in Table IX, automatic initialization of the tracker decreases the accuracy by 1.37%, 1.90%, and 0.97% (absolute) for eyelid, cheek, and lip corner features, respectively. The decrease in classification performance is maximum for the cheek region, because the cheek area has a smooth skin texture and consequently the tracking strongly relies on the initialized surrounding landmarks. Correct classification rate of using all regions (mid-level fusion, voting) is 87.82% with automatic landmarking, where manual initialization provides an accuracy of 89.84%.

Additionally, we evaluate the effect of automatic initialization of the tracker for different databases. As shown in Table X, automatic initialization reduces the correct classification rates by only 1.65% (absolute) on average. Therefore, since the decrease is not statistically significant ($p > 0.10$, using t-test analysis), automatically initialized tracking is used in the remainder of our experiments.

I. Comparison With Other Methods

We compare our method with the state-of-the-art smile classification systems proposed in the literature, namely, by Cohn and Schmidt [9], Dibeklioğlu *et al.* [25], and Pfister *et al.* [31]. To this end, we evaluate them on the UvA-NEMO database with the same experimental protocols, as well as on BBC, MMI and SPOS databases. [9] employs a linear discriminant classifier using duration, amplitude, and $\frac{\text{duration}}{\text{amplitude}}$ measures of smile onsets. [25] models eyelid movements during smiles to classify genuine and posed enjoyment smiles, where changes in eye aperture are described by distance-based and angular features. [31] models smile onsets using spatio-temporal CLBP-TOP (Completed LBP from Three Orthogonal Planes) features. Their method enables the use of infrared images in addition to natural texture images. Since infrared images are absent from UvA-NEMO, BBC, and MMI databases, results for this method are given by using only natural texture images.

All the methods have been implemented by us, since they are not publicly available. Only the CLBP-TOP feature extractor used in [31], has been provided by Pfister *et al.* The original implementation of the method proposed by Dibeklioğlu *et al.* [25] is used in our experiments. For a fair comparison, all methods

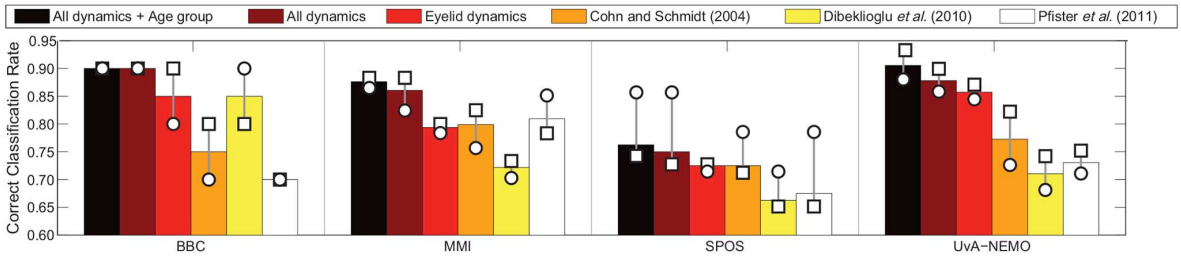

Fig. 13. Comparison of different methods on the BBC, MMI, SPOS, and UvA-NEMO databases. Bars show the mean accuracy of the related method, where square and circle markers indicate the classification rates of spontaneous and posed smiles, respectively.

are tested by using the PBVD tracker [32] with automatic initialization [33]. The proposed methods with age information enabled, and using solely eyelid features are also included in the comparison. To automatically estimate the required age information, the method proposed by [46] is used. Accuracies for all these methods are given in Fig. 13.

Results show that the proposed method outperforms the state-of-the-art methods. For the UvA-NEMO database, mid-level fusion with weighted SUM provides an accuracy of 87.82%, which is 10.56% (absolute) higher than the performance of the method proposed by Cohn and Schmidt [9]. Moreover, including labels of age groups in the feature set significantly ($p = 0.028$, using t-test analysis) improves the accuracy of our method and the resulting method reaches an accuracy of 90.56%. Using only the eyelid features decreases the correct classification rate by only 2.09% (absolute) in comparison to the mid-level fusion. This confirms the reliability of eyelid movements and the discriminative power of the proposed dynamical eyelid features to distinguish between types of smiles.

Our system with eyelid features has an 85.73% accuracy, significantly higher than that of [25] (71.05%), which uses only eyelid movement features without any temporal segmentation. This shows the importance of temporal segmentation. The accuracy of [31] (73.06%) is less than the accuracy of our method with only onset features, and shows that spatio-temporal features are not as reliable as dynamics.

Since the method in [9] relies on solely the onset features of lip corners, we have also tested our method with onset features of lip corners. We have obtained an accuracy of 80.73% compared to a 77.26% accuracy of [9]. We conclude that using automatically selected features from a large pool of informative features serves better than enabling a few carefully selected measures for this problem. Manually selected features may also show less generalization power across different (database-specific) recording conditions.

It is important to note that the proposed method uses solely onset features for the SPOS corpus, since it has only onset phases of smiles. Except for the SPOS corpus, we have observed that spontaneous smiles are generally classified better than posed ones for all methods. One possible explanation is that facial dynamics have more variance in posed smiles. Subsequently, the class boundaries of spontaneous smiles are more defined, and this leads to a higher accuracy.

When we analyze the change in the accuracy of our method with respect to different databases, it is seen that using only dynamics performs best on the BBC database. This is an expected result, since the variance of facial actions in both spontaneous and posed smiles of the BBC database is very limited. On the other hand, the highest accuracies for $All\ dynamics + Age\ group$ and *Eyelid dynamics* are achieved on the UvA-NEMO database. This can be based on the age annotations and the high frame rate (50 fps). Both the proposed methods and the competitors perform worst on SPOS corpus. This finding can be explained by the fact that the SPOS corpus only includes the onset phases of the smiles. Besides, lower frame rate of the recordings in SPOS (25 fps) can cause loss of some temporal information that results in lower accuracy.

VI. Discussion

In our experiments, the onset features of lip corners perform best for individual phases. This result is consistent with the findings of [9]. However, when onset, apex, and offset phases are fused, the eyelid movements are more descriptive than those of the cheeks and lip corners for enjoyment smile classification.

For the UvA-NEMO database, the best fusion scheme increases the correct classification rate by only 2.09% (absolute) with respect to the accuracy of eyelid features. This finding supports our motivation and confirms the discriminative power and the reliability of eyelid movements to classify enjoyment smiles. However, it is important to note that temporal segmentation of the smiles are performed by using lip corner movements, which means that additional information from the movements of lip corners is leveraged.

Using ANOVA, significant ($p < 0.001$, $\eta^2 > 0.05$) feature differences (of selected features) between *adults* and *young people* are obtained. For both spontaneous and posed smiles, the maximum and mean apertures of eyes are larger for *adults*. During the onset, both the amplitude of eye closure and closure speed are higher for *young people*. During the offset, amplitude and speed of eye opening are higher for *young people*. When we analyze the significance levels of the most selected features for smile classification, we conclude that the size of the eye aperture is smaller during spontaneous smiles. However, many subjects in the UvA-NEMO lower their eyelids also during posed smiles. This result is consistent with the findings in [18], which indicate that the D-marker can exist during both spontaneous and posed enjoyment smiles.

Another important finding is that the speed and acceleration of the eyelid movements are higher for posed smiles. As a result, since faster eyelid movements of *young people* cause confusion

with posed smiles, the classification accuracy with eyelid features is higher for *adults*. Similarly, features extracted from the cheek region perform better for *adults*, since cheek movements of *adults* are slower and more stationary. The duration of spontaneous smiles are longer than posed ones, but the lip corner movement for posed smiles is faster (also have higher acceleration). This improves the accuracy of the classification with lip corner features in favor of *young people*, since the lip corner movements of *young people* are significantly faster than *adults* during posed smiles.

Since eyelid and cheek features are reliable in *adults* as opposed to lip corners in *young people*, regional fusion in age-specific method decreases the accuracy compared to the performance of the generic method. On the other hand, including labels of estimated age groups in the feature set improves the accuracy by 2.27% (absolute) on the UvA-NEMO Smile Database in comparison to using only expression dynamics. To find out the reason behind that, the most selected dynamic features are analyzed using multivariate analysis of variance and significant ($p < 0.001$, $\eta^2 > 0.10$) feature differences between different ages are found. Our findings indicate that the dynamics of smile onsets are more affected by age than other smile phases. During the onset phase of smiles, the maximum speed and the maximum acceleration of both eye closure and lip corner movements significantly change among different ages. More detailed analysis of age related smile dynamics is given by [46], which uses the proposed features (facial dynamics) for age estimation.

Moreover, when analyzing the effect of gender on smile classification, the most significant ($p < 0.001$, $\eta^2 > 0.035$, using ANOVA) feature differences (of selected features) between *males* and *females* are obtained. Resulting findings show that maximum and mean apertures of eyes are significantly larger for *females* during apex, onset and offset phases. However, the rest of the dynamics do not differ significantly between genders. This may explain the accuracy decrease when the gender information is used in the system. We have also looked at gender-specific training, but did not obtain improved results. Such classifier specification reduces the number of training samples per classifier (effectively halving it for gender-specific classifiers in a gender balanced training set), and the improvements due to specification do not necessarily improve the final results.

Lastly, there is no significant symmetry difference (in terms of amplitude) between spontaneous and posed smiles as indicated by [29], [15]. We have also failed to find significant effects of symmetry between *young people* and *adults*, or between *males* and *females*.

VII. Conclusion

In this paper, we have provided an extensive discussion of facial expression dynamics for spontaneous and posed smile analysis. Based on a set of informative features extracted from a closely tracked facial image, an accurate smile classifier has been described.

Our results show that among facial regions, eyelid features are more relevant (compared to cheek and lip corner features) for smile analysis, but fusing all regions is useful. Similarly, the smile onset is the most informative phase of the smile, but adding apex and offset information is beneficial. For smile classification, our results suggest that mid-level fusion is more suitable compared to late (decision-level) fusion.

In this paper we have systematically evaluated how age information affects smile classification, and established that different face regions are differently affected by aging in terms of dynamics. Eyelid features are significantly more informative in adults, whereas cheek and lip corner features are more informative for young people. By designing age-specific classifiers, we can thus improve smile classification, even if the age is estimated automatically.

We have evaluated the proposed system using the UvA-NEMO Smile Database, which we have recently introduced. With 1240 samples of 400 subjects, this is the largest database in the literature for this task, and since the ages of subjects vary from 8 to 76 years, it allows a thorough investigation of age-related effects. Additionally, we have reported comparative evaluations on three smile datasets from the literature. Our proposed method is contrasted to three smile classification approaches from the literature over these four databases, and consistently outperforms them.

Acknowledgment

This research was part of Science Live, the innovative research program of science center NEMO that enables scientists to carry out real, publishable, peer-reviewed research using NEMO visitors as volunteers.

References

[1] P. Ekman and W. V. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco, CA, USA: Consulting Psychol. Press Inc., 1978.
[2] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[3] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 39–58, Jan. 2009.
[4] Z. Ambadar, J. F. Cohn, and L. I. Reed, "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *J. Nonverbal Behav.*, vol. 33, no. 1, pp. 17–34, 2009.
[5] P. Ekman, W. V. Friesen, and M. O'Sullivan, "Smiles when lying," *J. Personality Social Psychol.*, vol. 54, no. 3, pp. 414–420, 1988.
[6] P. Ekman, *Telling Lies: Cues to Deceit in the Marketplace, Politics, and Marriage*. New York, NY, USA: Norton, 1992.
[7] P. Ekman, J. C. Hager, and W. V. Friesen, "The symmetry of emotional and deliberate facial actions," *Psychophysiol.*, vol. 18, pp. 101–106, 1981.
[8] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *J. Nonverbal Behav.*, vol. 6, pp. 238–252, 1982.
[9] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 2, no. 2, pp. 121–132, 2004.
[10] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in *Proc. ECCV*, 2012, pp. 526–539.
[11] B. Duchenne, *The Mechanism of Human Facial Expression*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[12] P. Ekman, R. J. Davidson, and W. V. Friesen, "The Duchenne smile: Emotional expression and brain physiology II," *J. Personality Social Psychol.*, vol. 58, no. 2, pp. 342–353, 1990.
[13] P. Ekman, G. Roper, and J. C. Hager, "Deliberate facial movement," *Child Develop.*, vol. 51, pp. 886–891, 1980.
[14] P. Ekman, "Darwin, deception, and facial expression," *Ann. New York Academy Sci.*, vol. 1000, no. 1, pp. 205–221, 2003.

[15] K. L. Schmidt, S. Bhattacharya, and R. Denlinger, "Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises," *J. Nonverbal Behav.*, vol. 33, pp. 35–45, 2009.
[16] P. Gosselin, M. Perron, and M. Beaupré, "The voluntary control of facial action units in adults," *Emotion*, vol. 10, no. 2, p. 266, 2010.
[17] S. D. Gunnery, J. A. Hall, and M. A. Ruben, "The deliberate duchenne smile: Individual differences in expressive control," *J. Nonverbal Behav.*, vol. 37, no. 1, pp. 29–41, 2013.
[18] E. G. Krumhuber and A. S. R. Manstead, "Can Duchenne smiles be feigned? New evidence on felt and false smiles," *Emotion*, vol. 9, no. 6, pp. 807–820, 2009.
[19] V. Manera, M. D. Giudice, E. Grandi, and L. Colle, "Individual differences in the recognition of enjoyment smiles: No role for perceptual–attentional factors and autistic-like traits," *Frontiers Psychol.*, vol. 2, 2011.
[20] G. Littlewort-Ford, M. S. Bartlett, and J. R. Movellan, "Are your eyes smiling? Detecting genuine smiles with support vector machines and Gabor wavelets," in *Proc. Joint Symp. Neural Comput.*, 2001.
[21] M. Del Giudice and L. Colle, "Differences between children and adults in the recognition of enjoyment smiles," *Develop. Psychol.*, vol. 43, no. 3, p. 796, 2007.
[22] D. M. Shore and E. A. Heerey, "The value of genuine and polite smiles," *Emotion*, vol. 11, no. 1, p. 169, 2011.
[23] J. C. Borod, E. Koff, and B. White, "Facial asymmetry in posed and spontaneous expressions of emotion," *Brain Cognition*, vol. 2, no. 2, pp. 165–175, 1983.
[24] M. F. Valstar and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proc. ACM ICMI*, 2007, pp. 38–45.
[25] H. Dibeklioğlu, R. Valenti, A. A. Salah, and T. Gevers, "Eyes do not lie: Spontaneous versus posed smiles," in *Proc. ACM Multimedia*, 2010, pp. 703–706.
[26] L. Zhang, D. Tjondronegoro, and V. Chandran, "Geometry vs. appearance for discriminating between posed and spontaneous emotions," in *Proc. ICONIP*, 2011, pp. 431–440.
[27] M. He, S. Wang, Z. Liu, and X. Chen, "Analyses of the differences between posed and spontaneous facial expressions," in *Proc. ACII*, 2013, pp. 79–84.
[28] J. F. Cohn, L. I. Reed, T. Moriyama, J. Xiao, K. L. Schmidt, and Z. Ambadar, "Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, May 2004, pp. 129–135.
[29] L. K. Schmidt, Z. Ambadar, J. F. Cohn, and L. I. Reed, "Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling," *J. Nonverbal Behav.*, vol. 30, no. 1, pp. 37–52, 2006.
[30] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, "Spontaneous vs. posed facial behavior: Automatic analysis of brow actions," in *Proc. ACM ICMI*, 2006, pp. 162–170.
[31] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *Proc. ICCV Workshops*, 2011, pp. 868–875.
[32] H. Tao and T. Huang, "Explanation-based facial motion tracking using a piecewise Bézier volume deformation model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 1999, vol. 1, pp. 611–617.
[33] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "A statistical method for 2-D facial landmarking," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 844–858, Feb. 2012.
[34] T. W. Sederberg and S. R. Parry, "Free-form deformation of solid geometric models," *ACM SIGGRAPH Comput. Graphics*, vol. 20, no. 4, pp. 151–160, 1986.
[35] K. L. Schmidt, J. F. Cohn, and Y. Tian, "Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles," *Biological Psychol.*, vol. 65, no. 1, pp. 49–66, 2003.
[36] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[37] S. Petridis, B. Martinez, and M. Pantic, "The MAHNOB laughter database," *Image Vis. Comput.*, vol. 31, no. 2, pp. 186–202, 2013.
[38] M. F. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proc. LREC, Workshop Emotion*, 2010, pp. 65–70.
[39] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.
[40] M. Pantic and I. Patras, "Detecting facial actions and their temporal segments in nearly frontal-view face image sequences," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2005, vol. 4, pp. 3358–3363.
[41] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 28–43, Feb. 2012.
[42] S. Koelstra and M. Pantic, "Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, Sep. 2008, pp. 1–8.
[43] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Segment and recognize expression phase by fusion of motion area and neutral divergence features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, Mar. 2011, pp. 330–335.
[44] L. I. Kuncheva, "Fusion of continuous-valued outputs," in *Combining Pattern Classifiers: Methods and Algorithms*. New York, NY, USA: Wiley, 2004, pp. 151–188.
[45] L. G. Farkas, *Anthropometry of the Head and Face*. New York, NY, USA: Raven Press, 1994.
[46] H. Dibeklioğlu, T. Gevers, A. A. Salah, and R. Valenti, "A smile can reveal your age: Enabling facial dynamics in age estimation," in *Proc. ACM Multimedia*, 2012, pp. 209–218.
[47] N. Ramanathan, R. Chellappa, and S. Biswas, "Computational methods for modeling facial aging: A survey," *J. Vis. Lang. Comput.*, vol. 20, no. 3, pp. 131–144, 2009.
[48] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.
[49] G. Guo, R. Guo, and X. Li, "Facial expression recognition influenced by human aging," *IEEE Trans. Affective Comput.*, vol. 4, no. 3, pp. 291–298, Jul.-Sep. 2013.
[50] G. Guo and X. Wang, "A study on human age estimation under facial expression changes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun, 2012, pp. 2547–2553.
[51] C. Zhang and G. Guo, "Age estimation with expression changes using multiple aging subspaces," in *Proc. BTAS*, 2013, pp. 1–6.
[52] Y. Saatci and C. Town, "Cascaded classification of gender and facial expression using active appearance models," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, Apr. 2006, pp. 393–398.
[53] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTS: The extended M2VTS database," *AVBPA*, vol. 964, pp. 965–966, 1999.
[54] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Biometrics and Identity Management*, ser. Lecture Notes Comput. Sci.. Berlin, Germany: Springer, 2008, vol. 5372, pp. 47–56.
[55] A. Martinez and R. Benavente, "The AR Face Database," CVC Tech. Rep. 24, 1998.

Hamdi Dibeklioğlu (S'08–M'14) received the B.Sc. degree in computer engineering from Yeditepe University, Istanbul, Turkey, in 2006, the M.Sc. degree in computer engineering from Boğaziçi University, Istanbul, Turkey, in 2008, and the Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 2014.

He is currently a Post-Doctoral Researcher with the Pattern Recognition and Bioinformatics Group, Delft University of Technology, Delft, The Netherlands. He is also a Guest Researcher with the Intelligent Systems Lab Amsterdam, University of Amsterdam. His research interests include computer vision, pattern recognition, and automatic analysis of human behavior.

Dr. Dibeklioğlu received the Alper Atalay Second Best Student Paper Award at the IEEE Signal Processing and Communications Applications Conference in 2009. He was the recipient of the Best Presentation Award at the LongTerm Detection and Tracking Workshop of CVPR in 2014. He served on the local organization committee of the eNTERFACE Workshop on Multimodal Interfaces in 2007 and 2010.

Albert Ali Salah (M'08) received the Ph.D. degree in computer engineering from Boğaziçi University, Istanbul, Turkey.

Between 2007 and 2011 he worked at the CWI Institute and the Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands. He is currently an Assistant Professor with the Computer Engineering Department, Boğaziçi University, where he is also the Chair of the cognitive science program. He has authored or coauthored over 100 publications, including the book *Computer Analysis of Human Behavior* (Springer, 2011). His research interests include computer vision, multimodal interfaces, pattern recognition, and computer analysis of human behavior.

Dr. Salah received the inaugural EBF European Biometrics Research Award in 2006 for his work on facial feature localization. He was Co-Chair for the 6th eNTERFACE International Workshop on Multimodal Interfaces, the 14th ACM International Conference on Multimodal Interaction, and the 15th International Conference on Scientometrics and Informetrics. He initiated the International Workshop on Human Behavior Understanding in 2010 and was Co-Chair between 2010 and 2014. He served as a Guest Editor for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and *IEEE Pervasive Computing*, and as an Associate Editor of the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT. He is a member of the IEEE AMD Technical Committee Taskforce on Action and Perception, and the IEEE Biometrics Council.

Theo Gevers (M'02) is a Full Professor of Computer Vision with the University of Amsterdam (UvA), Amsterdam, The Netherlands, and a part-time Full Professor with the Computer Vision Center, Barcelona, Spain. He is a Founder of Sightcorp and 3DUniversum, spinoffs of the Intelligent Systems Laboratory, UvA. His main research interests are in the fundamentals of image understanding, 3-D object recognition, and color in computer vision.

Dr. Gevers is the Chair for various conferences and is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is a Program Committee Member for a number of conferences and an Invited Speaker at major conferences. He has given lectures at various major conferences, including the IEEE Conference on Computer Vision and Pattern Recognition, the International Conference on Pattern Recognition, SPIE, and the Computer Graphics, Imaging, and Vision Conference.