Combining Facial Dynamics With Appearance for Age Estimation

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

***Abstract*— Estimating the age of a human from the captured images of his/her face is a challenging problem. In general, the existing approaches to this problem use appearance features only. In this paper, we show that in addition to appearance information, facial dynamics can be leveraged in age estimation. We propose a method to extract and use dynamic features for age estimation, using a person's smile. Our approach is tested on a large, gender-balanced database with 400 subjects, with an age range between 8 and 76. In addition, we introduce a new database on posed disgust expressions with 324 subjects in the same age range, and evaluate the reliability of the proposed approach when used with another expression. State-of-the-art appearance-based age estimation methods from the literature are implemented as baseline. We demonstrate that for each of these methods, the addition of the proposed dynamic features results in statistically significant improvement. We further propose a novel hierarchical age estimation architecture based on adaptive age grouping. We test our approach extensively, including an exploration of spontaneous versus posed smile dynamics, and gender-specific age estimation. We show that using spontaneity information reduces the mean absolute error by up to 21%, advancing the state of the art for facial age estimation.**

***Index Terms*— Age estimation, age grouping, expression dynamics, smile, disgust, spontaneity.**

I. Introduction

AGE estimation from human faces is a challenging problem with applications in forensics, security, biometrics, electronic customer relationship management, entertainment and cosmetology [1]–[3]. Automatic age estimation can augment many computer applications in these domains, but it can also be used as a stand-alone tool, since humans are not universally successful in estimating age. The most frequently used measure of age estimation is the mean absolute error (MAE), and a recent crowd-sourcing study performed with frequently used aging databases show that humans have a MAE of 7.2–7.4 years for estimating the age of a person over 15, depending on the database conditions [4].

The main challenge of age estimation is the heterogeneity in facial feature changes due to aging for different humans. To determine facial changes associated with age is a hard problem, because they are related not only to gender and to genetic properties, but also to a number of external factors such as health, living conditions and weather exposure. Gender can play a role in the aging process as there are differences in aging patterns and features in males and females. Furthermore, facial cosmetics, surgical operations, the presence of scars, and even the presence of facial hair can be mitigating factors for age estimation.

Age estimation is an active topic today due to the growing necessity of including this information in real-world systems. This necessity comes from the fact that age is important to understand requirements or preferences in different aspects of the daily life of a person. Systems implementing age specific human computer interaction can cope with these aspects. Some examples are biometric systems that filter their database for the estimated age range of a subject, vending machines capable of denying some products such as alcohol or cigarettes to an underage customer, or advertisements in different automated environments (web pages, displays in stores, etc.) that can be personalized according to the age of the individual interacting with the system.

Automatic facial age estimation is affected by the traditional factors that make face analysis difficult in general. Unknown illumination conditions, non-frontal facial poses, and presence of facial expressions, are some issues that such systems need to deal with. Especially, facial expressions might negatively affect the accuracy of automated systems: When a person smiles, for instance, wrinkles are formed and these can be misleading when only the appearance cues are taken into account [5]. Similarly, sagging of the face in a sad expression can resemble the effects of aging.

The most important cues that are used in age classification are appearance-based, most notably the wrinkles formed on the face due to deformations in skin tissue. For this reason, current systems mainly focus on static appearance features of the face,

Manuscript received January 31, 2014; revised July 21, 2014; accepted January 29, 2015. Date of publication March 13, 2015; date of current version March 27, 2015. This work was supported in part by the Dutch National Program COMMIT and in part by Boğaziçi University, Istanbul, Turkey, under Project BAP-6531. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joseph P. Havlicek.

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

F. Alnajar is with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Amsterdam 1098 XH, The Netherlands (e-mail: f.alnajar@uva.nl).

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

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

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/TIP.2015.2412377

as it is the easiest way to obtain satisfactory results [2]. Hence, the dynamics of facial movement are largely ignored.

In this paper, instead of only considering static appearance features, we explore a novel set of dynamic features for age estimation. As movement features can be observed from facial expressions, the aim is to use dynamic features derived from these facial expressions for estimating the age. Since the smile is one of the most frequently used facial expressions, as well as the easiest emotional facial expression to pose voluntarily [6], we first focus on smiles and analyze the discrimination power of smile dynamics for age estimation. Once we verify that smile dynamics can improve discrimination, we validate the effectiveness of the proposed approach on a different facial expression.

There are a number of changes that happen on the face with aging, including loss of muscle tone, loss of underlying fat tissue, which reduces the smoothness of the face and creates wrinkles, receding gums (and sometimes, missing teeth), increased crows feet around the eyes, sunken eyes as a consequence of fat from eyelids settling into eye sockets, texture changes like blotches, dark spots, bone mass reduction causing lower jaw to reduce in size, and cartilage growth to lengthen the nose [7]–[9]. All these morphological changes alter the overall appearance of an expression on the face, but especially the loss of muscle tone directly affects the dynamics, along with the appearance. It is well known that the elastic fibers on the face show fraying and fragmentation at advancing age [10]. By leveraging the deformation features of the facial surface patches, age estimation with dynamic features may improve over systems that use solely appearance-based cues.

The main contribution of this paper is to show, on multiple expressions, that expression dynamics can be used to better estimate the age of a person. We propose a fully automatic age estimation framework, and show that it significantly outperforms the generic approach. We also introduce the high-resolution UVA-NEMO Disgust Database, which we make publicly available. We report our results with smile and disgust expressions, and make our experimental protocols available.

We extend our previous study [11] in many ways. Apart from a more in-depth treatment and extended literature, (1) we use 3D volume changes via surface patches instead of landmark movements, (2) we add frequency and facial asymmetry descriptors to the feature set, (3) we use a two-level adaptive classification scheme, (4) we evaluate four appearance features, (5) we systematically analyze gender-specific and spontaneity-specific effects of aging features, (6) we introduce an adaptive grouping procedure, (7) we introduce a new public database for disgust expression and report results on it.

The next section introduces related work in age estimation. Since there are comprehensive surveys in this area [2], [3], we focus on the most successful approaches, and the most recent work. Section III describes the proposed system of age estimation. In particular, we describe the detection and tracking of facial landmarks, the set of dynamic features, and the two-level classification scheme. Section IV describes the experimental protocol and the UvA-NEMO Smile Database, as well as introducing the new UVA-NEMO Disgust Database. Section V reports extensive comparative results. We analyze the contribution of appearance and dynamic features in detail, selecting four different state-of-the-art appearance-based approaches to serve as baselines. We test the influence of different facial regions in age estimation, augment the method by using gender-specific analysis, and study the effects of spontaneity in facial expressions. It is followed by a discussion in Section VI. Section VII concludes the paper.

II. Related Work

Several works propose to determine facial pattern changes and evolution associated with the aging process, both from psychological and biological points of view. These studies are mostly aimed at age synthesis, i.e. changing the appearance of a rendered face to show proper effects of aging. Some of these works are useful in the determination of appropriate facial features for age estimation. For instance, O'Toole *et al.* [12] use 3D models of faces to apply caricaturing processes in order to describe age variations between samples. Wu *et al.* [13] develop a system for the simulation of wrinkles and skin aging for facial animation. Suo *et al.* [14] present a model for face aging by analyzing it as a Markov process through a graph representing different age groups. Tiddeman *et al.* [15] also develop prototype models for face aging using texture information. In [16], a quantitative approach to face evolution of aging is presented.

The results of these studies show that the craniofacial development and skin texture are the most important features for age estimation. In fact, one of the first approaches for age estimation is proposed by Kwon and da Vitoria Lobo [17], where individual faces are classified into three age groups (baby, young and senior). This classification is performed using the theory of craniofacial development [18] and facial skin wrinkle analysis. Lanitis *et al.* [19] propose an age estimation method based on regression analysis of the *aging function*. During the training procedure, a quadratic function of facial features is fitted to each individual in the training set as his/her aging function. As for age estimation, they propose four approaches to determine the proper aging function for the unseen face image. The Weighted Person Specific (WPS) approach achieves the best performance in the experiments. This function, however, relies on profiles of the individual containing external information such as gender, health, living style, etc.

Image processing methods, including tools for subspace learning and dimensionality reduction, are also used to automatically estimate the age. In [20], faces are projected into manifolds by using subspace learning followed by a regression model to estimate the age. The aging pattern subspace (AGES) method [21] models a sequence of individually aging face images by learning a subspace representation. The age of a test face is determined by the projection in the subspace that can reconstruct the face image best. This model is later extended by the authors to model the nonlinear nature of human aging by considering learning of nonlinear subspaces, using a model called KAGES (Kernel AGing pattErn Subspace) [22]. Zhan *et al.* [23] propose an extended non-negative matrix

factorization method to learn a subspace representation, which could recover age information while eliminating variations caused by identity, expression, pose, etc. Chen *et al.* propose a method that employs pairwise age ranking based on subspace learning for age prediction [24]. In their approach, age ranks from unlabeled data are incorporated by semi-supervised learning. [25] applies age-oriented local regression using distance metric learning and dimensionality reduction.

Feature extraction is one of the key issues of automatic age estimation. In [26], Guo *et al.* introduce biologically-inspired aging features (BIF) for age estimation. These features are based on Gabor filter responses for different orientations and scales. Alnajar *et al.* propose intensity- and gradient-based features to adopt a learning-based encoding method for age estimation under unconstrained imaging conditions [27]. For each pixel, neighboring pixels are sampled in a ring-based pattern to form a low-level feature vector. Then, the features are encoded using a PCA-tree-based codebook. [28] models the completed local binary patterns (CLBP) using an SVM regressor. Initially, the method fine-tunes facial alignments in terms of facial shape and pose. The similarity transformation is based on local binary pattern distributions.

Aging patterns show significant differences in young and elderly people, and human performance in age estimation shows differences for these groups. It seems possible to break the age estimation problem into simpler subproblems by adopting different strategies for different age groups. In [29], fuzzy age labels (human annotations) are used in combination with the real age labels to train an age estimation system. Fuzzy age labels are defined as the upper and lower bounds of human estimation. Hybrid constraint supported vector regression is proposed to model both deterministic and fuzzy labels. In [4], a hierarchical age estimation is proposed. It classifies each facial component into one of four disjoint age groups using an SVM-based binary decision tree. For each age group, a separate SVM regressor is trained to fine-tune the age prediction. Then, outputs for different components using different features are fused to estimate the final age.

Age estimation and expression recognition are rarely coupled, although several systems rely on similar features and classification paradigms for both problems. In [30], an age estimation method is proposed to cope with significant expression changes, using correlation learning and discriminant mapping. However, this methodology requires both neutral and expressive facial images for the same subject, since it is based on the correlation between pairs of expressions. More recently, Zhang and Guo propose a weighted random subspace method to deal with expression changes by improving the discriminative power of the aging features [31].

Remarkably, the temporal dynamics of faces have been ignored in age estimation. Until [11], the precursor of the present work, the only study is by [32] in which Hadid proposes to use volume LBP (VLBP) features to describe spatio-temporal information in videos of talking faces and classifies the ages of the subjects into five groups (child, youth, adult, middle-age, and elderly). However, VLBP features alone are not powerful enough and the proposed system could not reach the accuracy of static image-based age estimation. Therefore, we propose to use facial dynamics and explore the potential for obtaining useful cues from facial expressions which have been unexplored so far.

III. Method

The aim of the proposed method is to estimate the age of subject by using a sequence of images that show the subject displaying a facial expression as input. To this end, we focus on the smile expression, since it is one of the most frequently shown facial expression. Additionally, disgust expression is considered to evaluate the reliability and generalizability of the approach.

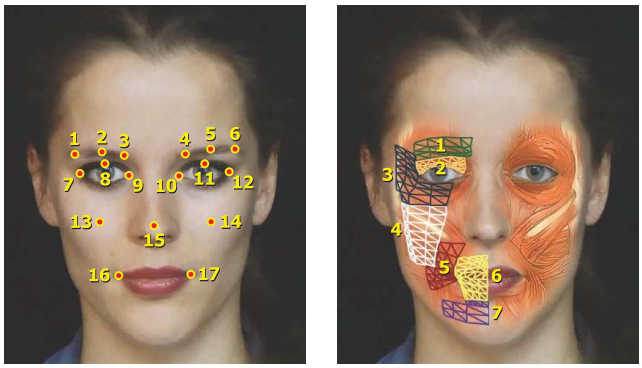
The proposed approach combines appearance features with facial expression dynamics. The method assumes that the input video starts with a moderately frontal face, and has the entire duration of a smile (or disgust) expression. These are typical assumptions of video-based expression recognition approaches.

The flow of the system is summarized as follows. Initially, a mesh model is fitted to face using 17 fiducial points, and tracked during the rest of the video. The surface deformations on different regions are computed using the tracked mesh points. Temporal phases (onset, apex, and offset) of the expression are estimated using the mean displacement signal of the lip corners. Then, dynamic features for each regional patch are extracted from each phase. Appearance features are extracted using the first frame of the onset phase, in which the face is neutral. After a feature selection procedure, the most informative dynamic features are selected and fused with appearance features to train Support Vector Machine (SVM) classifiers/regressors. In the rest of this section, we will outline the different components of our approach in detail.

A. Smile and Disgust Expressions

In this paper, we extract appearance and dynamic features from smile and disgust videos. In general, a smile can be modeled as the upward movement of the mouth corners, which corresponds to Action Unit 12 (AU12) in the facial action coding system (FACS) [33]. In terms of anatomy, the *zygomatic major* muscle contracts and raises the corners of the lips during a smile [34]. On the other hand, the disgust expression is the display of intense displeasure or condemnation that is shown by narrowing eyebrows, curling upper lip, and wrinkling nose. In terms of dynamics, smile and disgust expressions are composed of three non-overlapping phases; the onset, apex, and offset, respectively. Onset is the initial phase of a facial expression and it defines the duration from neutral to expressive state. Apex phase is the stable peak period of the expression between onset and offset. Likewise, the offset is the final phase from expressive to neutral state.

According to Ekman [6], there are many smiles, which are different in terms of their appearance and meaning. Ekman identified 18 of them (such as enjoyment, fear, miserable, embarrassment, listener response smiles) by describing the specific visual differences on the face and indicating the accompanying action units [6]. In this paper, we focus on enjoyment smiles for the detailed analysis, because they are



(a) (b)

Fig. 1. (a) Used facial feature points with their indices. (b) Regional surface patches (with their indices) and their underlying facial muscle structure. For simplicity, patches are shown on a single side of the face.

frequently shown and can easily be induced. Subsequently, we use posed and spontaneous smiles. To test whether the approach generalizes to other expressions, we use posed disgust expressions.

B. Facial Feature Tracking and Alignment

To analyze facial dynamics, surface deformations of seven facial regions (eyebrow, eyelid, eye-side, cheek, mouth-side, mouth, chin) are tracked in the videos [see Fig. 1(b)]. Patches for these regions are initialized in the first frame of the videos, using automatically detected 17 landmarks (corners and center of eyebrows, eye corners, center of upper eyelids, nose tip, and lip corners) for precise tracking and analysis [see Fig. 1(a)]. For automatic facial landmark detection, the method proposed by Dibeklioğlu *et al.* [35] is used. This method models Gabor wavelet features of a neighborhood of the 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 in which landmarks are initially detected on a coarse level and then fine-tuned for higher resolution. To track the facial features and pose, we use a piecewise Bézier volume deformation (PBVD) tracker, originally proposed by Tao and Huang [36].

The PBVD tracker employs a model-based approach. A 3D mesh model of the face [see Fig. 1(b)] 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 is defined as:

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

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:

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

After fitting the face model, facial feature points (as well as head motion) are tracked in 3D according to the movement and the deformations of the mesh. To measure 2D motion, template matching is used between frames at different resolutions. The estimated 2D image motion is modeled as a projection of the 3D movement onto the image plane. Then, the 3D movement is calculated using the projective motion of several points.

The tracked 3D coordinates of the facial feature points ℓ_i [see Fig. 1(a)] are used to align the faces in each frame. We estimate the 3D pose of the face, and normalize the face with respect to roll, yaw, and pitch rotations. Since three non-collinear 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 inner and outer eye corners and denoted by $c_1 = \frac{\ell_7+\ell_9}{2}$ and $c_2 = \frac{\ell_{10}+\ell_{12}}{2}$. 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:

$$\theta = \arccos \frac{U.\mathcal{N}_\mathcal{P}}{\|U\|\,\|\mathcal{N}_\mathcal{P}\|}, \quad \text{where } \mathcal{N} = \overrightarrow{\ell_{15}c_2} \times \overrightarrow{\ell_{15}c_1}. \quad (3)$$

In Equation 3, $\overrightarrow{\ell_{15}c_2}$ and $\overrightarrow{\ell_{15}c_1}$ denote the vectors from point ℓ_{15} 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 estimates the roll (θ_z) and yaw (θ_y) angles of the face with respect to the camera. However, the estimated pitch (θ_x) angle is subject-dependent, since it is relative to the constellation of the eye corners and the nose tip. If we assume that the face is approximately frontal in the first frame, then the actual pitch angles (θ'_x) are calculated by subtracting the initial value. Once the pose of the head is estimated, tracked points are normalized with respect to rotation, scale, and translation by:

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

where l_i is the aligned point and R_x, R_y, and R_z denote the 3D rotation matrices for the given angles. ρ denotes the Euclidean distance between the given points. On the normalized face, the middle point between eye centers is located at the origin and the inter-ocular distance (distance between eye centers) is set to 100 pixels.

C. Dynamic Features

To analyze the dynamics of facial deformations during an expression, we extract a set of dynamic features from seven different patches on the face [see Fig. 1(b)]. These patches are defined based on the underlying facial muscle structure, since the direction and the length of such muscles cause the visual variations of expressions [37].

When the tracked points are normalized, onset, apex, and offset phases of the smile and disgust expressions are detected, using the approach proposed by Schmidt *et al.* [38], by calculating the amplitude of the expression as the distance of the right lip corner to the lip center during the expression. Since the faces are normalized, the lip center is calculated only once in the first frame. Differently from [38], we estimate the

expression amplitude as the mean 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 mean amplitude signal of the lip corners in the frame t:

$$\mathcal{D}_{\text{lip}}(t) = \frac{\rho(\frac{l_{16}^1+l_{17}^1}{2}, l_{16}^t) + \rho(\frac{l_{16}^1+l_{17}^1}{2}, l_{17}^t)}{2\rho(l_{16}^1, l_{17}^1)}, \tag{5}$$

where l_i^t denotes the 3D location of the i^{th} point in frame t. This estimate is smoothed by the 4253H-twice method [39]. Then, 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.

To extract dynamic features from the given facial regions, deformation amplitude ($\mathcal{D}$) of the j^{th} patch at time t is estimated by:

$$\mathcal{D}_j(t) = \frac{\sum_{i=1}^{n_j} \lambda(j,i,t)}{\sum_{i=1}^{n_j} \lambda(j,i,1)}, \quad j = \{1, 2, \ldots, 7\}, \tag{6}$$

where n_j shows the number of meshes in patch j. $\lambda(j,i,t)$ denotes the area of the i^{th} triangular mesh of patch j at time t. Let p_1, p_2, and p_3 be the corner points of the related mesh, then its surface area is calculated by:

$$\lambda = \sqrt{\gamma(\gamma - \rho(p_1, p_2))(\gamma - \rho(p_1, p_3))(\gamma - \rho(p_2, p_3))}, \tag{7}$$

where

$$\gamma = \frac{\rho(p_1, p_2) + \rho(p_1, p_3) + \rho(p_2, p_3)}{2}. \tag{8}$$

Deformation amplitudes $\mathcal{D}_j$ are hereafter referred to as amplitude signals. As shown in Eq. 6, amplitude signals ($\mathcal{D}_j$) are normalized by the initial patch area (area in the first frame of the onset) for the sake of analysis. In addition to the amplitudes, speed $\mathcal{V}$ and acceleration $\mathcal{A}$ signals are computed by using the first and the second derivatives of the amplitudes, respectively:

$$\mathcal{V}(t) = \frac{d\mathcal{D}}{dt}, \tag{9}$$

$$\mathcal{A}(t) = \frac{d^2\mathcal{D}}{dt^2} = \frac{d\mathcal{V}}{dt}. \tag{10}$$

All the calculated amplitude signals are smoothed by the 4253H-twice method [39], and then split into three phases as onset, apex, and offset, which are previously defined using the amplitude signal $\mathcal{D}_{\text{lip}}$ of the lip corners.

A summary of the proposed dynamic features is given in Table I. Note that the defined features are extracted separately for each phase of the expression. As a result, we obtain three feature sets for each of the surface patches. Each phase is further divided into increasing ($^+$) and decreasing ($^-$) segments, for each feature set. This allows a more detailed analysis of the feature dynamics. Most of these features were originally proposed to analyze smile expressions [11], and a similar set has been employed for automatic kinship estimation through smile dynamics [40]. The present study demonstrates that they are also powerfully descriptive for the disgust expression.

TABLE I
DEFINITIONS OF THE EXTRACTED FEATURES

Feature	Definition
Frequency Components:	$[\ \psi(1)\ ,\ \psi(2)\ , \ldots, \psi(10)\ ]$
Duration:	$\left[\ \frac{\eta(\mathcal{D}^+)}{\omega}\ ,\ \frac{\eta(\mathcal{D}^-)}{\omega}\ ,\ \frac{\eta(\mathcal{D})}{\omega}\ \right]$
Duration Ratio:	$\left[\ \frac{\eta(\mathcal{D}^+)}{\eta(\mathcal{D})}\ ,\ \frac{\eta(\mathcal{D}^-)}{\eta(\mathcal{D})}\ \right]$
Maximum Amplitude:	$\max(\mathcal{D})$
Mean Amplitude:	$\left[\ \frac{\sum \mathcal{D}}{\eta(\mathcal{D})}\ ,\ \frac{\sum \mathcal{D}^+}{\eta(\mathcal{D}^+)}\ ,\ \frac{\sum \lvert\mathcal{D}^-\rvert}{\eta(\mathcal{D}^-)}\ \right]$
STD of Amplitude:	$\text{std}(\mathcal{D})$
Total Amplitude:	$\left[\ \sum \mathcal{D}^+\ ,\ \sum \lvert\mathcal{D}^-\rvert\ \right]$
Net Amplitude:	$\sum \mathcal{D}^+ - \sum \lvert\mathcal{D}^-\rvert$
Amplitude Ratio:	$\left[\ \frac{\sum \mathcal{D}^+}{\sum \mathcal{D}^+ + \sum \lvert\mathcal{D}^-\rvert}\ ,\ \frac{\sum \lvert\mathcal{D}^-\rvert}{\sum \mathcal{D}^+ + \sum \lvert\mathcal{D}^-\rvert}\ \right]$
Maximum Speed:	$\left[\ \max(\mathcal{V}^+)\ ,\ \max(\lvert\mathcal{V}^-\rvert)\ \right]$
Mean Speed:	$\left[\ \frac{\sum \mathcal{V}^+}{\eta(\mathcal{V}^+)}\ ,\ \frac{\sum \lvert\mathcal{V}^-\rvert}{\eta(\mathcal{V}^-)}\ \right]$
Maximum Acceleration:	$\left[\ \max(\mathcal{A}^+)\ ,\ \max(\lvert\mathcal{A}^-\rvert)\ \right]$
Mean Acceleration:	$\left[\ \frac{\sum \mathcal{A}^+}{\eta(\mathcal{A}^+)}\ ,\ \frac{\sum \lvert\mathcal{A}^-\rvert}{\eta(\mathcal{A}^-)}\ \right]$
Net Ampl., Duration Ratio:	$\frac{(\sum \mathcal{D}^+ - \sum \lvert\mathcal{D}^-\rvert)\omega}{\eta(\mathcal{D})}$
Left/Right Ampl. Difference:	$\frac{\lvert\sum \mathcal{D}_L - \sum \mathcal{D}_R\rvert}{\eta(\mathcal{D})}$

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 $\mathcal{D}$. η defines the length (number of frames) of a given signal, and ω is the frame rate of the video. $\mathcal{D}_L$ and $\mathcal{D}_R$ define the amplitudes for the left and right sides of the face, respectively. ψ denote the Discrete Cosine Transform (DCT) coefficients of $\mathcal{D}$ and computed by:

$$\psi(k) = \frac{1}{\varphi(k)} \sum_{t=1}^{\eta(\mathcal{D})} \mathcal{D}(t) \cos\left(\frac{\pi(2t-1)(k-1)}{2\eta(\mathcal{D})}\right), \tag{11}$$

where

$$\varphi(k) = \begin{cases} \sqrt{\eta(\mathcal{D})} : & k = 1 \\ \sqrt{\frac{\eta(\mathcal{D})}{2}} : & 2 \le k \le \eta(\mathcal{D}) \end{cases} \tag{12}$$

Since a low frequency signal can be reconstructed efficiently by using only a few DCT coefficients, we enable the first 10 DCT coefficients ($\psi(k)$, $k = \{1, 2, \ldots, 10\}$) of the amplitude signals in the feature set. As a result, for each face region, seven 35D feature vectors are generated by concatenating these features.

In some cases, features cannot be calculated. For example, if we extract features from the amplitude signal of the mouth

patch using the onset phase, the decreasing segments can be an empty set ($\eta\left(\mathcal{D}^{-}\right)=0$). For such exceptions, all the features describing the related segments are set to zero. This is done to have a generic feature vector format which has the same features for different phases of each face region.

D. Appearance Features

To describe the appearance of faces, we use four different state-of-the-art descriptors: namely, intensity-based encoded aging features, gradient-based encoded aging features, biologically-inspired aging features, and local binary patterns (LBP). Details of these appearance descriptors are given in this section.

Intensity-based (IEF) and gradient-based encoded aging features (GEF) are proposed by Alnajar *et al.* [27]. These features are based on a learning-based encoding. A discriminative low-level feature is computed for each pixel. Then, the features are encoded by a PCA-tree-based codebook [41]. The face is divided into patches and the codes in each patch are described by a histogram. Finally, the patch histograms are concatenated together to form the aging descriptor. Two versions of the descriptor are used based on low-level features: GEF based on gradient histogram (to capture wrinkle details) and IEF based on intensity sampling (to capture skin texture and fine wrinkle details). For IEF, the neighboring intensities are sampled around each pixel in a ring-based pattern. 25 intensity values are sampled at the circumferences of two rings with $r=1$ (8 values) and $r=2$ (16 values) including the central pixel value itself. To extract GEF, the gradient directions are computed in an 8×8 neighborhood of each pixel. The gradient orientations are binned to equally-spaced bins over $0°-360°$, where the gradient magnitudes are accumulated. As in [27], Gaussian derivatives are chosen for calculating the gradient and the number of bins equals eight.

Biologically-inspired aging features (BIF) are introduced by Guo *et al.* [26] for age estimation. The features are extracted by applying two-layer filters. In the first layer, BIF uses Gabor filter responses for different orientations and scales. In the second layer, it assembles the responses from the first layer in a local area (with the same directions and adjacent scales "band") to a single value using max or standard deviation functions. The authors adapt the descriptor from [42] by introducing the standard deviation operation in creating the second layer and making the number of bands and orientations adaptive to the data. In our experiments, for sake of simplicity, 16 orientations and eight bands are computed to build the descriptor.

The original local binary patterns operator, which is proposed by Ojala *et al.* [43], takes the intensity value of the center pixel as threshold to convert the neighborhood pixels to a binary code. Computed binary codes describe the ordered pattern of the center pixel. This procedure is repeated for each pixel on the image and the histogram of the resultant 256 labels can then be used as a texture descriptor. In [44], Ojala *et al.* show that a large number of the local binary patterns contain at most two bitwise transitions from 0 to 1 or 1 to 0, which is called a uniform pattern. Therefore, during the computation of the histograms, the size of the feature vector can be significantly reduced by assigning different bins for each of the 58 uniform patterns and one bin for the rest. Uniform local binary patterns are used in experiments, and are hereafter referred to as LBP. Eight neighborhood pixels (on a circle with a radius of 1 pixel) are used to extract the LBP features.

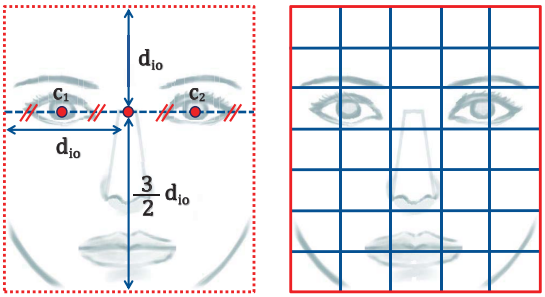


Fig. 2. (a) Cropping of a face image, and (b) the defined 7×5 blocks to extract appearance features.

Since the onset of a facial expression starts with a neutral face, the first frame of the previously detected onset phase is selected to extract the appearance features. On the selected frame, the roll rotation of the face is estimated and normalized using the eye centers c_1 and c_2. Then, the face is resized and cropped as shown in Fig. 2(a). The inter-ocular distance d_{io} is set to 50 pixels to normalize the scale and cropping. As a result, each normalized face image has a resolution of 125×100 pixels. After the preprocessing step, appearance features (IEF, GEF, BIF, and LBP) are computed. IEF, GEF, and LBP descriptors are extracted from 7×5 non-overlapping (equally-sized) blocks [see Fig. 2(b)]. For all descriptors, the dimensionality of the appearance feature vectors is reduced by Principal Component Analysis (PCA) so as to retain 99.99% of the variance.

E. Feature Selection and Classification

Estimating the age of a person by using a generic classifier/regressor is an inherently challenging problem, since many factors influence the age for different age groups (mainly shape in early ages and appearance in later ages [3]) and the learning-based predictor should capture all these details from the training data to produce a correct age estimation. One solution for this problem is dividing the prediction of the age into two phases: The first one predicts the age group. Next, a second fine-tuned age prediction model is learned to estimate the exact age.

In the two-level age prediction, the sample is first classified into an age group (first-level prediction). Later, another predictor will place the sample in its exact age (second-level prediction). In [11], the age groups are determined in a uniform way ($8-14$, $15-17$, $18-21$, $22-28$, $29-35$, $36-54$, $55-76$). However, problems may arise when boundary ages between two adjacent groups are not distinctive (i.e. the aging features are similar). In such cases, the first-level prediction is more prone to go wrong which is likely to propagate the error to the second level prediction. To overcome this issue, we propose a method that computes the ages which are

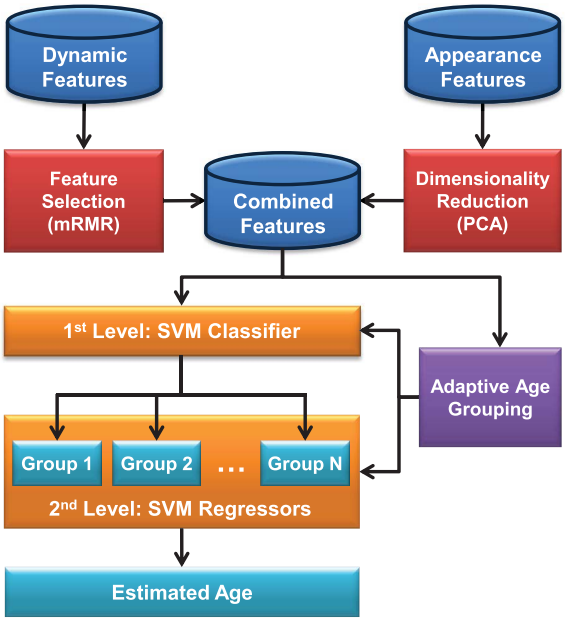


Fig. 3. Two-level age estimation architecture using both appearance and dynamic features.

dissimilar with their neighbors. The whole age range is divided into groups in such a way that the boundary between each two adjacent groups is discriminant. To this end, the average cosine similarity $\mathcal{S}$ between each age a and its $2q$ neighbors $h = \{a-q, a-q+1, \ldots, a+q\} - \{a\}$ is computed by:

$$\mathcal{S}_a = \frac{1}{2qn_a} \sum_{i=1}^{2q} \sum_{j=1}^{n_a} \sum_{k=1}^{n_{h_i}} \frac{d_a^j . d_{h_i}^k}{n_{h_i} \left\| d_a^j \right\| \left\| d_{h_i}^k \right\|}, \tag{13}$$

where d_a^j denotes the feature vector of age a's j^{th} sample. n_a denotes the number of samples for age a. After smoothing $\mathcal{S}$, age a is set to be a group boundary if $\forall h_i, \mathcal{S}_{h_i} > \mathcal{S}_a$. Minimum and maximum age values in the whole range are also set as boundaries. Each boundary age is included in the same group with its most similar adjacent neighbor. The number of neighborhood levels $q \geq 2$ is selected automatically on the validation data.

In the given two level architecture (see Fig. 3), we use Support Vector Machine classifiers and regressors for age estimation. In the first level, one-vs-all SVM classifiers are used to classify the age of a subject into automatically defined age groups. Then, the age of the subject is fine-tuned using an SVM regressor which is specifically trained for the related age group. For an improved estimation, the regressor of each age group is trained with an age interval of -10 to $+10$ years of group boundaries. Then, the results are limited by the age range (if the estimated age is less/more than the group boundaries, it is set to the minimum/maximum age of the group). The resulting estimation of the age is given as an integer with a 1 year resolution.

As described in Section III-C, we extract three 35D dynamic feature vectors for each face region. To deal with feature redundancy, we use the Min-Redundancy Max-Relevance (mRMR) algorithm to select the discriminative dynamic features [45]. mRMR is an incremental method minimizing the redundancy while selecting the most relevant information as follows:

$$\max_{f_j \in F - S_{m-1}} \left[I\left(f_j, c\right) - \frac{1}{m-1} \sum_{f_i \in S_{m-1}} I\left(f_j, f_i\right) \right], \tag{14}$$

where I shows the mutual information function and c indicates the target class. F and S_{m-1} denote the feature set, and the set of $m-1$ features, respectively. Then, all the selected dynamic features are concatenated with the appearance features (which are extracted from the first frame of the expression onset and reduced by PCA) to train the system (see Fig. 3). Minimum classification error on a separate validation set is used to determine the most discriminative dynamic features. Similarly, to optimize the SVM configuration, different kernels (linear, polynomial, and radial basis function) with different parameters (degree of polynomial kernel) are tested on the validation set and the configuration with the minimum validation error is selected. The test partition of the dataset is not used for parameter optimization.

IV. Experimental Settings

A. UvA-NEMO Smile Database

The UvA-NEMO Smile Database[1] has been collected to analyze the change in dynamics of smiles for different ages [46]. Data collection was carried out in the Science Center NEMO (Amsterdam) [47] as part of Science Live, an innovative research programme. NEMO visitors are the volunteers for the data collection. The database and its evaluation protocols are made available to the research community.

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 are recorded with a resolution of 1920×1080 pixels at a rate of 50 frames per second under controlled illumination conditions. Additionally, a color chart is present on the background of the videos for illumination and color normalization. Sample frames from the database are shown in Fig. 4.

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. 43 subjects do not have spontaneous smiles and 32 subjects have no posed smile samples. Age and gender distributions of the subjects in the database are given in Fig. 5(a).

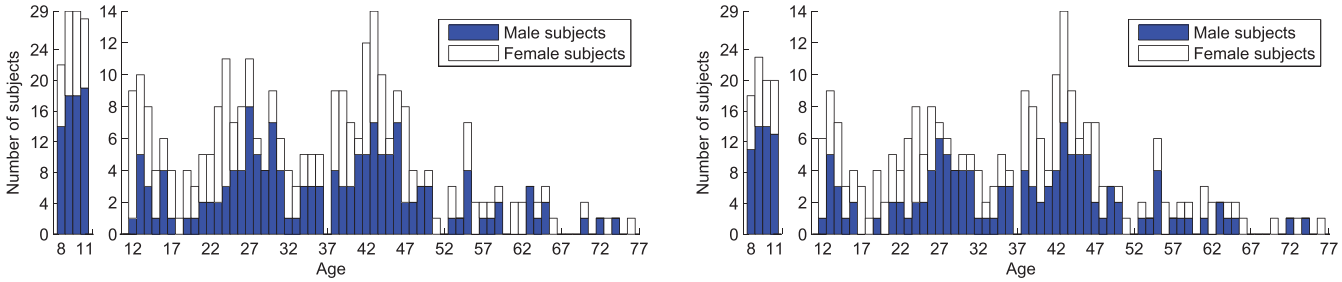
For posed smiles, each subject was asked to pose a smile as realistically as possible, sometimes after being shown the proper way in a sample video. Short, funny video segments are used to elicit spontaneous smiles. Approximately five minutes of recordings are made per subject, and genuine smiles are segmented.

For each subject, a balanced number of spontaneous and posed smiles are selected and annotated by seeking consensus of two trained annotators. Each segment starts and ends with neutral or near-neutral expressions.

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

(a) (b) (c) (d)

Fig. 4. Sample frames from the UvA-NEMO Smile and the UvA-NEMO Disgust Databases: (a) Showing neutral face, (b) posed enjoyment smile, (c) spontaneous enjoyment smile, and (d) disgust expression.



(a) (b)

Fig. 5. Age and gender distributions of the subjects in (a) the UvA-NEMO Smile and (b) the UvA-NEMO Disgust Databases.

B. UvA-NEMO Disgust Database

To show the applicability of the proposed approach on other facial expressions, we introduce the UvA-NEMO Disgust Database[1] in this paper. This database is composed of posed (deliberate) disgust expressions, and recorded during the collection of the UvA-NEMO Smile Database using the same recording/illumination setup. Sample frames from the database are shown in Fig. 4.

Each subject was asked to pose a disgust expression as realistically as possible, sometimes after being shown a sample video. For each subject, one or two posed disgust expressions were selected and annotated by seeking consensus of two trained annotators. Each segment starts and ends with neutral or near-neutral expressions. The resulting database has 518 deliberate disgust videos from 324 subjects (152 female, 172 male). 313 of 324 subjects are also included in the 400 subjects of the UvA-NEMO Smile Database. The ages of subjects vary from 8 to 76 years. Age and gender distributions of the subjects in the database are given in Fig. 5(b). The database and its evaluation protocols are made available to the research community.

C. Settings

To evaluate our system and assess the reliability of facial expression dynamics and facial appearance information for age estimation problem, we first use the UvA-NEMO Smile Database of 400 subjects. We then show the effectiveness of the proposed approach for disgust expression using the UvA-NEMO Disgust Database. In our experiments, the two-level classification/regression system is used as described in Section III-E. The optimum number of selected dynamic features, adaptive age ranges, and kernels of the SVM classifiers/regressors are determined on a separate validation partition. To this end, a two level 10-fold cross-validation scheme is used. 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. Candidate settings for these parameters are set as follows: Neighborhood level $q = \{2, 3, 4\}$ for adaptive age grouping; candidate kernels for the SVMs are linear, polynomial, and radial basis function. In polynomial and radial basis function kernels, the gamma parameter is set as $1/d$, where d is the number of features.

There is no subject overlap between folds in either database. We initialize the tracking by automatically annotated facial landmarks. For automatic facial landmark detection, we use the system proposed by Dibeklioğlu *et al.* [35]. The mean localization error for the related landmarks [corners and center of eyebrows, eye corners, center of upper eyelids, nose tip, lip corners; see Fig. 1(a)] is 3.84% of the inter-ocular distance to the actual location of the landmarks. Correlation coefficients between the extracted amplitude signals with manual and automatic initializations ranged between 0.93 and 1.

V. Experiments

In this section, we present the results of our experiments on exact age estimation. First, we evaluate the accuracy of the proposed system when only facial dynamics are used, either

TABLE II
EFFECT OF USING DIFFERENT FACIAL REGIONS WITH AND WITHOUT FEATURE SELECTION ON THE UVA-NEMO SMILE DATABASE

Regions	MAE (years)	
	without Feat. Selection	with Feat. Selection
1: Eyebrow	15.34 (±10.59)	13.32 (±9.63)
2: Eyelid	15.87 (±11.38)	13.50 (±10.21)
3: Eye-sides	14.74 (±10.15)	12.93 (±9.52)
4: Cheek	13.88 (±10.44)	12.14 (±9.35)
5: Mouth-sides	14.98 (±12.36)	13.27 (±11.42)
6: Mouth	15.74 (±13.73)	14.15 (±12.11)
7: Chin	28.70 (±29.70)	24.42 (±26.29)
1–7: All	12.04 (±9.81)	10.81 (±8.85)

individually for each facial region, or taken together, on smile expressions. We compare these results with the combined use of appearance and dynamics. We then test the influence of gender and expression spontaneity on the accuracy of the system using the combined features. Finally, we report age estimation results using disgust expression dynamics.

A. Dynamics

Since the proposed dynamic features are extracted from the deformations of seven different surface patches, we analyze the individual discrimination power of these deformations and their combination for age estimation. Furthermore, to assess the reliability of the feature selection step, performance of using automatically selected (most) informative dynamic features and the use of all features without any selection are compared. The resulting *mean absolute error* (MAE) is given in Table II.

As shown in Table II, the feature selection increases the accuracy by approximately 13% (relative) on average, while reducing the dimensionality of the feature space. Since the efficacy of the feature selection step is confirmed by these results, it is used in the remainder of our experiments. By analyzing the regional results with feature selection, it can be derived that the dynamics of cheek's surface deformations are the most reliable features, with an MAE of 12.14 (±9.35) years. Deformation dynamics on the sides of the eyes follow closely with an MAE of 12.93 (±9.52) years. The chin region provides an MAE of only 24.42 (±26.29) because of its stationary surface characteristic. By combining the dynamic features of different facial regions, the MAE of the age estimation is decreased to 10.81 (±8.85) years.

B. Dynamics Versus Appearance

The aim of this work is to improve the accuracy of age estimation by combining facial appearance with expression dynamics. However, it is also important to show the discriminative power of facial expression dynamics and appearance, individually. For this purpose, the individual and combined uses of these features are evaluated.

As shown in Table III, combining dynamics with appearance features significantly improves the age estimation accuracy in comparison to the individual use of dynamic and appearance features ($p < 0.0015$ for GEF, BIF and LBP; $p < 0.01$ for IEF). It is clear that the use of only facial dynamics is not sufficient for accurate age estimation. The MAE of using dynamic features is 10.81 (±8.85) years, where the MAEs for different facial appearance descriptors range from 4.80 (±4.77) to 5.78 (±6.15) years. However, by combining dynamic and appearance features, the proposed system is able to achieve the best results. The combination of dynamics and IEF provides the highest accuracy with an MAE of 4.33 (±4.03). The cumulative distribution of the MAEs for individual and combined features are shown in Fig. 6.

TABLE III
MEAN ABSOLUTE ERRORS FOR DYNAMICS, APPEARANCE, AND COMBINED FEATURES ON THE UVA-NEMO SMILE DATABASE

Features	MAE (years)	
	without *Dynamics*	with *Dynamics*
Appearance: None	N/A	10.81 (±8.85)
Appearance: IEF	4.80 (±4.77)	4.33 (±4.03)
Appearance: GEF	5.48 (±5.57)	4.82 (±4.89)
Appearance: BIF	5.78 (±6.15)	5.03 (±5.10)
Appearance: LBP	5.46 (±5.58)	4.77 (±4.66)

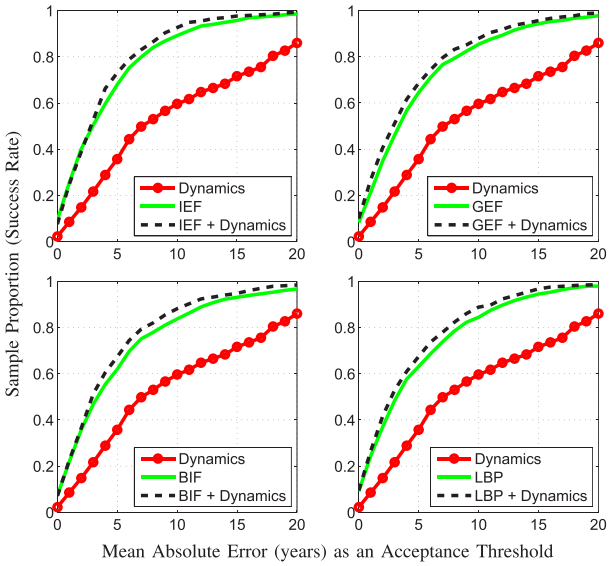


Fig. 6. Cumulative distribution of the mean absolute error for different features on the UvA-NEMO Smile Database.

The UvA-NEMO Database has uniform D65 illumination that does not specifically highlight wrinkles. For images with direct illumination, we observe that wrinkles become much more pronounced, and gradient based descriptors perform better than intensity based features.

C. Assessment of Adaptive Age Grouping

We now test the use of a two-level classification/regression strategy. We evaluate the performance of using three different classification/regression approaches, namely direct regression (no grouping), classifying age groups into bins of 10-years and

TABLE IV
PERFORMANCE OF ADAPTIVE AGE GROUPING, GROUPING INTO BINS OF 10 YEARS AND THE REGRESSION WITHOUT GROUPING (NONE) ON THE UVA-NEMO SMILE DATABASE

Features	MAE (years)		
	None	10-years	Adaptive
IEF + Dynamics	5.00 (±4.25)	4.40 (±4.11)	4.33 (±4.03)
GEF + Dynamics	5.63 (±4.86)	4.97 (±5.07)	4.82 (±4.89)
BIF + Dynamics	5.12 (±4.91)	5.94 (±5.24)	5.03 (±5.10)
LBP + Dynamics	5.29 (±4.36)	4.83 (±4.60)	4.77 (±4.66)

TABLE V
COMPARISON OF THE GENDER-SPECIFIC METHOD WITH THE GENERAL METHOD FOR AGE ESTIMATION ON THE UVA-NEMO SMILE DATABASE

Features	MAE (years)	
	Gender-specific	General
IEF + Dynamics	4.25 (±3.95)	4.33 (±4.03)
GEF + Dynamics	4.67 (±4.71)	4.82 (±4.89)
BIF + Dynamics	4.91 (±4.88)	5.03 (±5.10)
LBP + Dynamics	4.58 (±4.47)	4.77 (±4.66)

into adaptive bins (based on training) before group-specific regression. The resulting MAEs of each method for different feature combinations and their cumulative distributions are shown in Table IV and in Fig. 7, respectively.

The results show that the adaptive grouping outperforms other approaches for all features. 10-years grouping follows it, and provides a more accurate estimation in comparison to that of direct regression in most cases.

D. Effect of Gender

To assess the effect of gender on the accuracy of age estimation using the combined features, a gender-specific age estimation approach is implemented. In the gender-specific method, different classifiers/regressors are trained and tested for both males and females, separately. For this method, we assume that the gender labels of all samples are given correctly. The MAEs for both gender-specific and the general approach are given in Table V.

Our results show that the gender-specific training decreases the overall MAE in comparison the MAE of general-training. The MAE of the gender-specific approach for different features range from 4.91 (±4.88) to 4.25 (±3.95) years. Although the improvement is not statistically significant, the gender-specific training provides a 3% MAE enhancement (relative) on average.

In particular, the improvement for males is more than that of females. The cumulative distribution of the MAE for general and gender-specific methods are shown in Fig. 8.

E. Effect of Expression Spontaneity

To assess the effect of expression spontaneity on the accuracy of using combined features, a spontaneity-specific age estimation system is implemented. For this purpose, separate classifiers/regressors are trained for spontaneous and posed smiles. Spontaneity of smiles is classified using the system of [46]. This system uses similar expression dynamics to distinguish between spontaneous and posed smiles. Correct classification of the system is 87.02% on the UvA-NEMO Smile Database.

TABLE VI
COMPARISON OF THE SPONTANEITY-SPECIFIC METHOD WITH THE GENERAL METHOD FOR AGE ESTIMATION ON THE UVA-NEMO SMILE DATABASE

Features	MAE (years)	
	Spontaneity-specific	General
IEF + Dynamics	4.00 (±3.74)	4.33 (±4.03)
GEF + Dynamics	4.40 (±4.44)	4.82 (±4.89)
BIF + Dynamics	4.59 (±4.59)	5.03 (±5.10)
LBP + Dynamics	4.38 (±4.23)	4.77 (±4.66)

As shown in Table VI, the MAE of the spontaneity-specific approach ranges from 4.00 (±3.74) to 4.59 (±4.59), therefore improving the accuracy by 8% (on average) with respect to the general approach. This means a statistically significant ($p < 0.04$) improvement. Spontaneity-specific training decreases the MAE for both posed and spontaneous smiles. Since the automatically detected neutral faces are used to extract the appearance features for both approaches, accuracy improvements by performing spontaneity-specific training indicates the differences between spontaneous and posed smiles in terms of expression dynamics. The cumulative distribution of the MAE for general and spontaneity-specific methods are shown in Fig. 8.

F. Effect of Temporal Phases

An expression video from onset to offset contains a lot of frames. The system we have proposed gives a decision when the expression is completed, i.e. at the end of the offset phase. However, it may be necessary to give a decision while the expression is in progression. To understand how partial information would fare, we implement a version of the proposed method. Spontaneity specific approach (with automatic spontaneity detection) is used with adaptive age grouping in this experiment.

Since the order of the temporal phases during a facial expression is fixed, the online system starts classification in the onset mode, where appearance features are combined with onset dynamics. When the apex is reached, it uses both the onset and the apex in addition to appearance. In the final stage, dynamics of all three phases are used together with appearance features. For these three modes, separate classifiers are trained.

The performance of the implemented system for the UvA-NEMO Smile Database is given in Table VII. The results show that while all phases contribute to the accuracy, the highest improvement rates are provided by combining onset dynamics with appearance features. Including onset dynamics in the feature set decreases the MAE by 11.76% (relative)

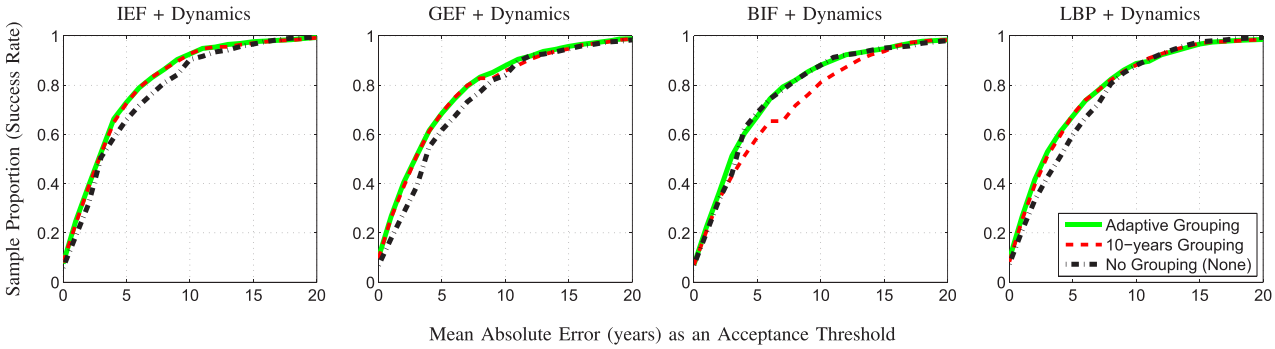


Fig. 7. Cumulative distribution of the mean absolute error for different grouping strategies on the UvA-NEMO Smile Database.

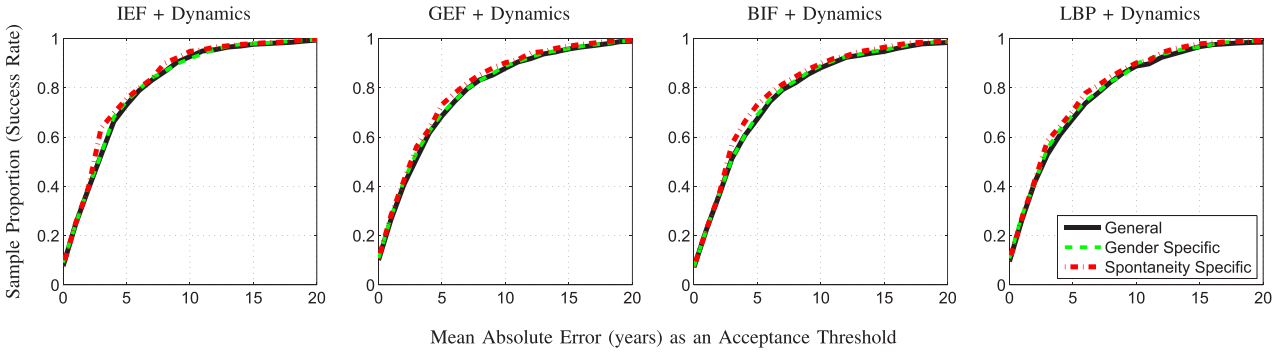


Fig. 8. Cumulative distribution of the mean absolute error for general, gender-specific, and spontaneity-specific methods on the UvA-NEMO Smile Database.

TABLE VII
EFFECT OF USING DYNAMICS OF DIFFERENT TEMPORAL PHASES FOR AGE ESTIMATION ON THE UVA-NEMO SMILE DATABASE

Features	MAE (years)			
	w/o Dynamics	+Onset	+Onset to Apex	+Onset to Offset
IEF	4.80 (±4.77)	4.36 (±4.24)	4.23 (±4.09)	4.00 (±3.74)
GEF	5.48 (±5.57)	4.92 (±5.02)	4.75 (±4.71)	4.40 (±4.44)
BIF	5.78 (±6.15)	5.08 (±5.36)	4.87 (±4.95)	4.59 (±4.59)
LBP	5.46 (±5.58)	4.83 (±4.98)	4.66 (±4.67)	4.38 (±4.23)

on average. Including apex and offset phases in the analysis, increases the accuracy further.

G. Comparison to Other Methods

To the best of our knowledge, this is the first study using facial expression dynamics (such as speed, acceleration, amplitude, etc.) for age estimation. Except for the recent work by [32], none of the previous studies in the literature focus on using temporal information for age estimation.

In [32], Hadid proposes to use spatio-temporal information to classify the ages of the subjects into five groups (child: 0 to 9 years old; youth: 10 to 19; adult: 20 to 39; middle-age: 40 to 59 and elderly: above 60). They use volume LBP (VLBP) features with a tree of four SVM classifiers. VLBP features are extracted from different overlapping face blocks. Then the AdaBoost learning algorithm is used to determine the optimal size and locations of the local rectangular prisms, and select the most discriminative VLBP features for classification, automatically. To evaluate the system, 2000 videos of about 300 frames are randomly segmented from a set of video sequences mainly showing talking faces (collected from the Internet). Additionally, an appearance-based (static) system is implemented for comparison. This baseline method classifies each frame in a video, individually, using LBP features with SVM classifiers. Majority voting is used to fuse the classification results of each frame. Hadid reports that the static image (appearance) based approach provides 77.4% correct classification, where the performance of the spatio-temporal approach reaches only 69.2%.

VLBP is a straightforward extension of the original LBP operator to describe dynamic textures (image sequences) [48]. VLBP enables the use of temporal space (T), models the face sequence as a volume, and the neighborhood of each pixel is defined in 3D space. In contrast, LBP uses only X and Y dimensions of a single image. Then, the histograms of VLBP are used as features. In [48], Zhao and Pietikäinen have proposed to extract LBP histograms from Three Orthogonal Planes (LBP-TOP) XY, XT, and YT, individually, and concatenate them as a single feature vector.

To compare our system with related approaches using smiles, we implement three baseline methods: (1) VLBP-based spatio-temporal approach, (2) spatio-temporal approach using LBP-TOP features, and (3) appearance-based approach which classifies the first and the last frame of a smile onset (a neutral and an expressive face, respectively) using appearance features, individually, and fuses the estimations by mean operator. All methods use the same classification/regression architecture as our method. Tests are performed on the

TABLE VIII
MEAN ABSOLUTE ERROR ON THE UVA-NEMO SMILE DATABASE FOR DIFFERENT METHODS

	Method	MAE (years) for Different Age Ranges								
		0–9	10–19	20–29	30–39	40–49	50-59	60–69	70–79	All
Dynamics	Deformation	4.85	8.72	12.22	13.06	13.53	11.55	14.13	17.82	10.81 (±8.85)
	Displacement [11]	5.42	9.67	11.98	14.53	12.77	15.42	20.57	20.35	11.54 (±11.49)
Appearance	IEF, Fusion	2.73	3.28	4.68	5.36	5.38	9.09	12.97	14.65	4.86 (±4.54)
	GEF, Fusion	2.89	3.17	5.00	6.07	5.38	10.97	16.90	17.53	5.24 (±5.38)
	BIF, Fusion	3.92	3.73	4.90	5.61	5.86	12.30	17.33	18.24	5.63 (±5.79)
	LBP, Fusion	3.42	4.02	5.13	6.63	5.60	9.06	10.43	13.41	5.37 (±5.28)
Combined, Spon. Spec.	IEF + Dynamics	2.25	2.50	3.74	4.59	4.34	8.20	11.07	13.35	4.00 (±3.74)
	GEF + Dynamics	1.40	2.29	3.99	5.17	5.37	10.17	15.00	16.06	4.40 (±4.44)
	BIF + Dynamics	2.68	3.21	4.23	5.09	4.62	9.44	13.17	14.12	4.59 (±4.59)
	LBP + Dynamics	1.53	2.68	3.95	5.31	5.31	9.33	11.83	13.94	4.38 (±4.23)
Spatio-temporal	VLBP [32]	10.69	12.95	15.99	18.54	18.43	16.58	23.80	26.59	15.70 (±12.40)
	LBP-TOP	9.71	11.01	14.19	15.88	16.75	15.29	19.70	23.71	13.83 (±10.97)
	Number of Samples	158	333	215	171	250	66	30	17	1240

UvA-NEMO Smile Database. For a fair comparison, all of the compared methods use automatically annotated facial landmarks to initialize the tracking (for face alignment and feature extraction), and 7×5 non-overlapping blocks on the face to compute the histograms. To generate histograms, uniform patterns are used for LBP-TOP and LBP. The neighborhood size is set to eight for LBP and LBP-TOP, and two for VLBP. Time interval for the volumetric approaches is set to three frames. Zhao and Pietikäinen [48] show that these neighborhood and time interval parameters perform well for facial expression classification. To provide comparable smile durations for spatio-temporal descriptors, each smile phase (onset, apex, and offset) is temporally interpolated to 25 frames using bicubic interpolation. The dimensionality of IEF, GEF, BIF, LBP, VLBP, and LBP-TOP features is reduced by Principal Component Analysis (PCA) so as to retain 99.99% of the variance.

As shown in Table VIII, the combination of dynamic and appearance features provides the most accurate results. The spontaneity-specific method that combines dynamics and IEF achieves the minimum MAE of 4.00 (±3.74) years. Note that combining appearance with expression dynamics provides more accurate age estimation than using neutral and expressive frames in a video and averaging the results.

Spatio-temporal methods can only reach a mean accuracy of 15.70 (±12.40) and 13.83 (±10.97) years with VLBP and LBP-TOP features, respectively. By the sole use of proposed dynamic features, the system is significantly more accurate than when it uses the spatio-temporal features ($p < 0.001$). Finally, we also compare the deformation-based dynamic features that proposed in this paper (first row of Table VIII) with the displacement dynamics (of eyelids, cheeks, and lip corners) introduced in our previous study [11] (second row of Table VIII), and show that surface deformation dynamics perform better.

TABLE IX
AVERAGE TIME REQUIREMENTS OF DIFFERENT FEATURE EXTRACTION MODULES

Module	Duration (seconds per frame)
Initial landmarking (four points: eye corners)	0.4323
Initial landmarking (17 points)	1.1204
Tracking	0.0411
Normalization of the landmarks	0.0017
Cropping/alignment of the face image	0.0234
IEF feature extraction	1.3228
GEF feature extraction	3.6268
BIF feature extraction	38.3390
LBP feature extraction	0.0042
Dynamics extraction (per signal)	0.0752

H. Computational Load

In this section, we report average time requirements of different feature extraction modules. All the modules except the Bézier volume tracker, are composed of non-optimized MATLAB code or C++/compiled MATLAB code. Speed tests are conducted on an Intel i7-3687U, 2.1GHz (dual core) processor with 16GBs of RAM.

Average time requirements for each module used for different features are given in Table IX. For a smile of 3.2 seconds (160 frames), facial expression dynamics can be extracted in 8.0436 seconds (based on landmarking of 17 points, tracking, normalization of landmarks, and dynamics extraction). Similarly, dynamics of a 0.6 seconds (30 frames) onset phase can be extracted in 2.4796 seconds. Extraction of IEF, GEF, BIF, and LBP features requires 1.7785, 4.0825, 38.7947, 0.4599 seconds per frame, respectively (based on landmarking of four points, cropping/alignment of the face image, and feature extraction).

TABLE X
MEAN ABSOLUTE ERRORS FOR DYNAMICS, APPEARANCE, AND COMBINED FEATURES ON THE UVA-NEMO DISGUST DATABASE

Features	MAE (years)	
	without *Dynamics*	with *Dynamics*
Appearance: None	N/A	10.32(±7.97)
Appearance: IEF	5.04 (±4.90)	4.21 (±4.07)
Appearance: GEF	5.50 (±5.67)	4.56 (±4.74)
Appearance: BIF	7.24 (±8.46)	6.09 (±7.08)
Appearance: LBP	5.29 (±5.05)	4.38 (±4.18)

Newer generation of trackers use local binary features for accurate and very fast face alignment, achieving speeds around 3000 fps on desktop computers [49]. Consequently, commercial systems will be able to perform face tracking and expression analysis in real time. Our results show that age estimation will also benefit from the availability of accurate and computationally cheap dynamic information.

I. Application to Disgust Expression

To evaluate the effectiveness of the proposed approach on a different facial expression, we conduct experiments on disgust expression. To this end, the UvA-NEMO Disgust Database is used. Adaptive age grouping is enabled in these experiments. Here, appearance features are extracted from the first, neutral frames of the onset of the disgust videos.

As shown in Table X, similar to the results on smiles, combining dynamics of disgust expression with appearance features significantly improves the age estimation accuracy in comparison to the individual use of dynamic and appearance features ($p < 0.05$). MAE improvement over the accuracy of appearance features ranges from 0.83 to 1.15 years. These improvement rates are higher than those given in Table III for smile videos. One reason is that all the disgust expressions are deliberate, causing the system to act like a spontaneity-specific system and thus providing better modeling.

J. Classification of Age Ranges

Based on different application requirements, many studies report automatic age estimation results as classification of age groups [50]–[52]. Since facial dynamics are much more informative for large age differences, we conduct a set of experiments to show the usefulness of dynamic features in classifying age groups using smile and disgust expressions. Two different set of age groups are evaluated in our experiments: (a) 7 age groups of 10 years (8–17, 18–27, …, 68–77) as in [11], and (b) 5 age groups (8–12, 13–19, 20–36, 37–65, 66+) as in [50]. Our system is trained for these two different sets of age ranges. Since the UvA-NEMO Smile Database includes both spontaneous and posed smiles, spontaneity specific system is used for smiles by employing the automatic spontaneity detection (for smiles) as proposed in [46]. The general classification approach is used for disgust expression, since the UvA-NEMO Disgust Database has only posed disgust expressions.

TABLE XI
CLASSIFICATION ACCURACY OF AGE RANGES FOR APPEARANCE, AND COMBINED FEATURES ON THE UVA-NEMO SMILE AND THE UVA-NEMO DISGUST DATABASES

Expr.	Feature	Classification Accuracy (%)			
		7 Groups		5 Groups	
Smile	IEF	67.50	75.73	76.13	88.55
	GEF	65.48	74.84	74.11	87.58
	BIF	60.56	72.34	71.94	84.76
	LBP	65.40	74.35	70.24	83.63
Disgust	IEF	64.29	76.06	79.34	90.93
	GEF	64.09	75.68	75.68	86.68
	BIF	56.95	69.50	64.67	78.76
	LBP	60.62	71.81	78.96	89.58
		without *Dynamics*	with *Dynamics*	without *Dynamics*	with *Dynamics*

As shown in Table XI, combining dynamics with appearance features significantly ($p < 0.01$) improves the classification accuracy of age groups. When smile videos are used, mean accuracy improvement for 7 and 5 age groups are 9.58%, and 13.03% (absolute), respectively. When disgust videos are used, mean accuracy improvement for 7 and 5 age groups are 11.78%, and 11.83% (absolute), respectively. In comparison to the improvement on exact age estimation, these results display a more visible enhancement. This is based on higher reliability of expression dynamics for classifying subjects with large age differences. When the estimation of exact age is considered, the use of the expression dynamics improves the first level classification accuracy in a visible way, but in the second level (regression), while the exact age within the classified group is being determined, dynamics fall short of fine-level estimation, and results mainly rely on appearance features. As a result, expression dynamics are much more reliable and discriminative for classifying age groups.

VI. DISCUSSION

In our experiments, we show that deformation dynamics of cheek's during smiles, perform best for individual regions. Additionally, fusion of all regions (with a feature selection step) improves the accuracy of the cheek dynamics by 10.96%. For dynamic features, using feature selection increases the accuracy approximately by 13% on average, as well as reducing feature dimensionality. This finding indicates that there is a significant amount of noise or confusing information in surface deformation dynamics.

Our results show that the individual use of the facial expression dynamics is not sufficient to obtain an accurate age estimation system. However, accuracy of using solely appearance features of a neutral face (automatically detected as the first frame of the onset phase) is significantly outperformed ($p < 0.0015$ for GEF, BIF and LBP; $p < 0.01$ for IEF) by enabling the surface deformation dynamics of smile expression. Moreover, the use of combined features

outperforms all the baseline methods tested in this study. These results confirm the importance of facial expression dynamics. We also show that the deformation-based dynamics outperform the displacement dynamics (of eyelids, cheeks, and lip corners) introduced in our previous study [11].

To obtain the most informative dynamic features for age estimation, we use the frequently selected descriptors (in feature selection procedure). Significant ($p < 0.001$, $\eta^2 > 0.15$) differences of these features between different ages are investigated using multivariate analysis of variance (MANOVA). Majority of the frequently selected features are extracted from onset and offset of smiles. Additionally, the differences of these features among different ages display lower significance level (p) in comparison to the apex features. Such findings indicate that the deformation dynamics of smile onsets and offsets are more discriminative than those of apex phase for age estimation. During the onset phase of smiles, the mean speed and the mean amplitude of deformation decrease ($\mathcal{D}^-$) on eyelids significantly change among different ages. During the apex phase, the maximum and mean amplitude of deformation on the mouth region are significantly different for different ages. When the offset features are analyzed, it is shown that the maximum and net amplitude of deformation on the mouth region significantly change. Additionally, the second frequency component ($\psi(2)$) of deformation amplitude on the mouth region significantly differs among ages. Note that $\psi(2)$ denotes the lowest frequency coefficient of the deformation amplitude, since $\psi(1)$ is always the DC-component ($\frac{\sum \mathcal{D}}{\sqrt{\eta(\mathcal{D})}}$). So it can be inferred that during smile offsets, the rough shape of the mouth deformation amplitude is an informative feature for age estimation.

Then, we have analyzed the significant differences of these features between spontaneous and posed smiles using the t-test. Our results show that the mean speed and the amplitude of deformation decrease ($\mathcal{D}^-$) on eyelids are significantly higher ($p < 0.005$) for posed smiles during the smile onsets. During the offset phase, on the mouth region, the second frequency component of deformation is lower ($p = 0.002$) for spontaneous smiles, whereas the net amplitude of deformation is ($p < 0.001$) significantly higher. Similarly, the t-test analysis is repeated for male and female differences. The results indicate that during smile apexes, the maximum and mean amplitude of the deformation of the mouth region is significantly lower for males ($p < 0.001$). During the offset phase, on the mouth region, the second frequency component and the maximum amplitude of deformation is lower ($p < 0.001$) for males, whereas the net amplitude of the deformation is ($p < 0.001$) significantly higher. These findings can explain the higher accuracy of the spontaneity- and gender-specific systems. The reader is referred to [53] for further analysis of smile dynamics.

Experimental results show that spatio-temporal approaches (using smiles) based on VLBP and LBP-TOP are not efficient for age estimation. Even the individual use of our dynamic features outperforms these methods significantly. Spatio-temporal features describe the change of facial appearance in time, but our proposed approach models the appearance on a single neutral image (which is automatically selected as the first frame of the onset phase) and adds the surface deformation dynamics of the facial expression (such as amplitude, speed, acceleration, etc) on it. As a result, the proposed system (using spontaneity-specific approach) is significantly ($p < 0.001$) more accurate than all the competitor methods used in our experiments. When we evaluate the performance of combining proposed features with appearance for disgust expression, similar to our results on smiles, age estimation accuracy is significantly improved.

Our additional experiments on the classification of age ranges demonstrate that the facial expression dynamics are much more reliable for group classification tasks. This is due to higher discrimination power of expression dynamics for classifying subjects with large age differences. However, dynamics are not discriminative enough for discriminating similar ages. This finding can be explained by the large variation of expression dynamics within a narrow age range.

VII. Conclusions

Our study shows that dynamic facial features obtained by analyzing a frequently occurring facial expression improves appearance-based age estimation. While appearance is more informative than facial dynamics, it is affected by many external factors, like make-up, scars, and wrinkles resulting from exposure to harsh weather conditions. Such factors do not concern facial dynamics. Consequently, dynamics are sufficiently uncorrelated with appearance to allow fusion approaches for age estimation.

In our previous work, we have assessed a range of dynamical features in an exploratory fashion, and have shown that if landmark movements are employed, eyelid dynamics are the most revealing in terms of age estimation, followed by lip corners and cheeks [11]. The present work improves these results by using surface area features (instead of landmark movements) for characterizing 3D facial dynamics. We also introduce in this paper a two-level classifier, where the age range for each classifier is adaptively selected in the first level. We test four different features for appearance to show that the improvement by dynamic features is consistent across representations, and we also introduce an appearance fusion baseline. We study gender effects systematically, to conclude that the improvement due to gender-specific models is nor significant. We show that spontaneous and posed smiles have different and distinct dynamics, and spontaneity-specific age estimation significantly outperforms the general approach. Finally, we demonstrate that the method we propose is usable with other expressions, and report results on the new UvA-NEMO Disgust Database we introduce in this paper. Subsequently, this is the most extensive dynamic age evaluation study to this date in the literature.

References

[1] A. M. Albert, K. Ricanek, Jr., and E. Patterson, "A review of the literature on the aging adult skull and face: Implications for forensic science research and applications," *Forensic Sci. Int.*, vol. 172, no. 1, pp. 1–9, 2007.

[2] 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.

[3] 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.

[4] H. Han, C. Otto, and A. K. Jain, "Age estimation from face images: Human vs. machine performance," in *Proc. Int. Conf. BTAS*, Jun. 2013, pp. 1–8.

[5] M. Lucassen, T. Gevers, and H. Dibeklioğlu, "The effect of smile and illumination color on age estimation from faces," *Perception*, vol. 41, ECVP Abstract Supplement, p. 87, 2012. [Online]. Available: http://www.perceptionweb.com/abstract.cgi?id=v120501

[6] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York, NY, USA: Norton, 1992.

[7] S. E. Brodie, "Aging and disorders of the eye," in *Brocklehurst's Textbook of Geriatric Medicine and Gerontology*, H. M. Fillit, K. Rockwood, and K. Woodhouse, Eds., 7th ed. Philadelphia, PA, USA: Saunders, 2010, ch. 96.

[8] K. L. Minaker, "Common clinical sequelae of aging," in *Goldman's Cecil Medicine*, L. Goldman and A. I. Schafer, Eds., 24th ed. Philadelphia, PA, USA: Saunders, 2011, ch. 24.

[9] S. J. Salasche, "Anatomy," in *Flaps and Grafts in Dermatologic Surgery*, 1st ed., T. E. Rohrer, J. L. Cook, T. H. Nguyen, and J. R. Mellette Jr., Eds. Philadelphia, PA, USA: Saunders Elsevier, 2007, ch. 1.

[10] R. Sanders, "Torsional elasticity of human skin in vivo," *Pflügers Arch. Eur. J. Physiol.*, vol. 342, no. 3, pp. 255–260, 1973.

[11] 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 20th Int. Conf. Multimedia*, 2012, pp. 209–218.

[12] A. J. O'Toole, T. Vetter, H. Volz, and E. M. Salter, "Three-dimensional caricatures of human heads: Distinctiveness and the perception of facial age," *Perception*, vol. 26, no. 6, pp. 719–732, 1997.

[13] Y. Wu, N. M. Thalmann, and D. Thalmann, "A dynamic wrinkle model in facial animation and skin ageing," *J. Vis. Comput. Animation*, vol. 6, no. 4, pp. 195–205, 1995.

[14] J. Suo, S.-C. Zhu, S. Shan, and X. Chen, "A compositional and dynamic model for face aging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 385–401, Mar. 2010.

[15] B. Tiddeman, M. Burt, and D. Perrett, "Prototyping and transforming facial textures for perception research," *IEEE Comput. Graph. Appl.*, vol. 21, no. 5, pp. 42–50, Sep./Oct. 2001.

[16] M. Ortega, L. Brodo, M. Bicego, and M. Tistarelli, "On the quantitative estimation of short-term aging in human faces," in *Proc. 15th ICIAP*, 2009, pp. 575–584.

[17] Y. H. Kwon and N. da Vitoria Lobo, "Age classification from facial images," *Comput. Vis. Image Understand.*, vol. 74, no. 1, pp. 1–21, 1999.

[18] T. R. Alley, *Social and Applied Aspects of Perceiving Faces*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

[19] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

[20] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[21] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 307–316.

[22] X. Geng, K. Smith-Miles, and Z.-H. Zhou, "Facial age estimation by nonlinear aging pattern subspace," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 721–724.

[23] C. Zhan, W. Li, and P. Ogunbona, "Age estimation based on extended non-negative matrix factorization," in *Proc. IEEE 13th Int. Workshop Multimedia Signal Process.*, Oct. 2011, pp. 1–6.

[24] Y.-L. Chen and C.-T. Hsu, "Subspace learning for facial age estimation via pairwise age ranking," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2164–2176, Dec. 2013.

[25] W.-L. Chao, J.-Z. Liu, and J.-J. Ding, "Facial age estimation based on label-sensitive learning and age-oriented regression," *Pattern Recognit.*, vol. 46, no. 3, pp. 628–641, 2013.

[26] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE CVPR*, Jun. 2009, pp. 112–119.

[27] F. Alnajar, C. Shan, T. Gevers, and J.-M. Geusebroek, "Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions," *Image Vis. Comput.*, vol. 30, no. 12, pp. 946–953, 2012.

[28] J. Ylioinas, A. Hadid, X. Hong, and M. Pietikäinen, "Age estimation using local binary pattern kernel density estimate," in *Proc. 17th ICIAP*, 2013, pp. 141–150.

[29] J. Liu, Y. Ma, L. Duan, F. Wang, and Y. Liu, "Hybrid constraint SVR for facial age estimation," *Signal Process.*, vol. 94, pp. 576–582, Jan. 2014.

[30] G. Guo and X. Wang, "A study on human age estimation under facial expression changes," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2547–2553.

[31] C. Zhang and G. Guo, "Age estimation with expression changes using multiple aging subspaces," in *Proc. IEEE 6th Int. Conf. BTAS*, Sep./Oct. 2013, pp. 1–6.

[32] A. Hadid, "Analyzing facial behavioral features from videos," in *Proc. 2nd Int. Workshop Human Behavior Understand.*, 2011, pp. 52–61.

[33] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco, CA, USA: Consulting Psychologists Press, 1978.

[34] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *J. Nonverbal Behavior*, vol. 6, no. 4, pp. 238–252, 1982.

[35] 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.

[36] H. Tao and T. S. Huang, "Explanation-based facial motion tracking using a piecewise Bézier volume deformation model," in *Proc. IEEE Comput. Soc. Conf. CVPR*, vol. 1. Jun. 1999, pp. 611–617.

[37] L. R. Rubin, "The anatomy of a smile: Its importance in the treatment of facial paralysis," *Plastic Reconstructive Surgery*, vol. 53, no. 4, pp. 384–387, 1974.

[38] 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.

[39] P. F. Velleman, "Definition and comparison of robust nonlinear data smoothing algorithms," *J. Amer. Statist. Assoc.*, vol. 75, no. 371, pp. 609–615, 1980.

[40] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Like father, like son: Facial expression dynamics for kinship verification," in *Proc. IEEE ICCV*, Dec. 2013, pp. 1497–1504.

[41] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma, "Learning the structure of manifolds using random projections," in *Advances in Neural Information Processing Systems 20*. Red Hook, NY, USA: Curran Associates, 2007, pp. 473–480.

[42] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[43] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[44] T. Ojala, M. Pietikäinen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[45] 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.

[46] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in *Proc. ECCV*, 2012, pp. 525–538.

[47] *Science Center NEMO*. [Online]. Available: http://www.e-nemo.nl/, accessed Jan. 1, 2014.

[48] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[49] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 1685–1692.

[50] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 256–263.

[51] K. Ramesha, K. B. Raja, K. R. Venugopal, and L. M. Patnaik, "Feature extraction based face recognition, gender and age classification," *Int. J. Comput. Sci. Eng.*, vol. 1, no. 1S, pp. 14–23, 2010.

[52] K. Ueki, T. Hayashida, and T. Kobayashi, "Subspace-based age-group classification using facial images under various lighting conditions," in *Proc. 7th Int. Conf. AFGR*, Apr. 2006, pp. 43–48.

[53] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Recognition of genuine smiles," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 279–294, Mar. 2015.

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, 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 a recipient of the Best Presentation Award at the Long Term Detection and Tracking Workshop of IEEE Computer Society Conference on Computer Vision and Pattern Recognition in 2014. He served on the Local Organization Committee of the eNTERFACE Workshop on Multimodal Interfaces, in 2007 and 2010, respectively.

Fares Alnajar (S'13) received the M.Sc. (*cum laude*) degree in artificial intelligence from the University of Amsterdam, Amsterdam, The Netherlands.

He is currently pursuing the Ph.D. degree with the Intelligent Systems Lab Amsterdam, University of Amsterdam. He was an Intern with the Video and Image Processing Group, Philips Research Eindhoven. His research focuses on human behavior analysis.

Mr. Alnajar was a recipient of the Amsterdam Merit Scholarship, a Full Merit Scholarship for outstanding students coming from outside the European Union.

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

He was with the CWI Institute and the Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands, from 2007 to 2011. He is currently an Assistant Professor with the Department of Computer Engineering, Boğaziçi University, where he is also the Chair of the Cognitive Science Program. He has authored or co-authored over 100 publications, including the book entitled *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 is a member of the IEEE AMD Technical Committee Taskforce on Action and Perception, and the IEEE Biometrics Council. He received the inaugural EBF European Biometrics Research Award for his work on facial feature localization in 2006. He was the 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 the Co-Chair from 2010 and 2014. He served as a Guest Editor for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the *IEEE Pervasive Computing*, and as an Associate Editor of the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT.

Theo Gevers (M'01) is currently 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 the 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 a Program Committee Member for a number of conferences and an Invited Speaker at major conferences. He is the Chair for various conferences and is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He has given lectures at various major conferences, including the IEEE Conference on Computer Vision and Pattern Recognition, the International Conference on Pattern Recognition, International Society for Optics and Photonics, and the Computer Graphics, Imaging, and Vision Conference.