A Smile Can Reveal Your Age: Enabling Facial Dynamics in Age Estimation

Hamdi Dibeklioğlu[‡], Theo Gevers[‡], Albert Ali Salah[§], and Roberto Valenti[‡]
[‡]Intelligent Systems Lab Amsterdam, University of Amsterdam, Amsterdam, The Netherlands
[§]Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
h.dibeklioglu@uva.nl, th.gevers@uva.nl, salah@boun.edu.tr, r.valenti@uva.nl

ABSTRACT

Estimation of a person's age from the facial image has many applications, ranging from biometrics and access control to cosmetics and entertainment. Many image-based methods have been proposed for this problem. In this paper, we propose a method for the use of dynamic features in age estimation, and show that 1) the temporal dynamics of facial features can be used to improve image-based age estimation; 2) considered alone, static image-based features are more accurate than dynamic features. We have collected and annotated an extensive database of face videos from 400 subjects with an age range between 8 and 76, which allows us to extensively analyze the relevant aspects of the problem. The proposed system, which fuses facial appearance and expression dynamics, performs with a mean absolute error of 4.81 (± 4.87) years. This represents a significant improvement of accuracy in comparison to the sole use of appearance-based features.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video analysis; H.1.2 [**User/Machine Systems**]: Human factors, Human information processing

General Terms

Human Factors, Algorithms, Experimentation

Keywords

Age estimation, Facial dynamics

1. INTRODUCTION

The human face offers a rich and heterogeneous amount of data which have an important role in the communication process between individuals. By analyzing faces, humans can derive different useful attributes such as identity, gender, age, emotion, etc. Nowadays, with the pervasive presence of cameras and processing devices, intelligent systems dedicated to human-computer interaction tasks are also expected to be able to analyze and process the same information. A lot of effort has already been made in the literature on the estimation of each of these human-specific attributes.

In particular, age estimation is a very 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 daily life of a person. Systems implementing age specific human computer interaction (ASHCI) can cope with these aspects. Some examples are 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.

Age estimation from human faces is a challenging problem with a host of applications in forensics, security, biometrics, electronic customer relationship management, entertainment and cosmetology [1, 8, 22]. The main challenge is the huge heterogeneity in facial feature changes due to aging for different humans. Being able to determine the facial changes associated with age is a hard problem, because they are related not only to gender and to genetic properties, but to a number of external factors such as health, living conditions and weather exposure. Furthermore, gender can also play a role in the aging process, as there are differences in the aging patterns and features in males and females.

Facial expressions might negatively affect the accuracy of automated systems: When a person smiles, for instance, wrinkles are formed and these can be misleading if only the appearance cues are taken into account. Similarly, sagging of the face in a sad expression can resemble the effects of aging. Previous work in the literature tried to cope with this problem, generally by dividing the problem into different sub classes (i.e. one per facial expression), yielding mixed results.

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 the skin tissue. For this reason, current state of the art systems in the literature mainly focus on static appearance features of the face, as it is the easiest way to obtain accurate results. Hence, the dynamics of facial expressions are largely ignored.

Instead of only considering static appearance features for age estimation, we explore a novel set of features for age estimation in this paper, namely facial feature dynamics. As movement features can be observed during facial expressions, the aim is to use 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 [5], we focus on smiles and analyze the discrimination power of smile dynamics for age estimation. Our hypothesis is that aging effects the speed with

MM'12, October 29–November 2, 2012, Nara, Japan.

which facial expressions are formed on the face. It is well known that the elastic fibres on the face show fraying and fragmentation at advancing age [23]. By leveraging the movement features of points of interest on the face, we show that we can improve age estimation over systems that use solely appearance-based cues.

2. RELATED WORK

Several works have proposed methods 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 have been useful in the determination of appropriate facial features for age estimation. For instance, O'Toole *et al.* [19] used 3D models of faces to apply caricaturing processes in order to describe age variations between samples. Wu *et al.* [30] developed a system for simulation of wrinkles and skin aging for facial animation. Suo *et al.* [26] presented a model for face aging processing by analyzing it as a Markov process in a graph, representing different age groups. Tiddeman *et al.* [28] also developed prototype models for aging face images using texture information. In [18], a quantitative approach to face evolution due to aging is presented.

The results of these studies determined that cranio-facial development and skin texture are the most important features for age estimation. In fact, one of the first approaches for age estimation was proposed by Kwon and Lobo [13], where individual faces were classified into three age groups (baby, young and senior). This classification was performed using the theory of cranio-facial development [2] and facial skin wrinkle analysis. Lanitis *et al.* [14] proposed an age estimation method based on regression analysis of the *aging function*. During the training procedure, a quadratic function of facial features is fitted for each individual in the training set as his/her aging function. As for age estimation, they proposed four approaches to determine the suitable aging function for the unseen face image, among which the Weighted Person Specific (WPS) approach achieved the best performance in the experiments. This function, however, relied on profiles of the individual containing external information such as gender, health, living style, etc. In [11] a method for age estimation is presented, where faces are projected into manifolds by using subspace learning and then a regression model is applied to estimate the age. The aging pattern subspace (AGES) method [10] 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 the best. This model was later extended by the authors to model the nonlinear nature of human aging by considering learning of nonlinear subspaces, using the model called KAGES (Kernel AGing pattErn Subspace) [9].

Recently, Zhan *et al.* has proposed an extended non-negative matrix factorization method to learn a subspace representation, which could recovers age information while eliminating variations caused by identity, expression, pose, etc. [31]. In [12], Hadid has proposed using volume LBP (VLBP) features to describe spatiotemporal information in videos of talking faces and classify the ages of the subjects into five groups (child, youth, adult, middle-age, and elderly). This is the only study in the literature that uses temporal information for age estimation. However, VLBP features alone are not powerful enoughi and the proposed system could not even reach the accuracy of static image-based age estimation.

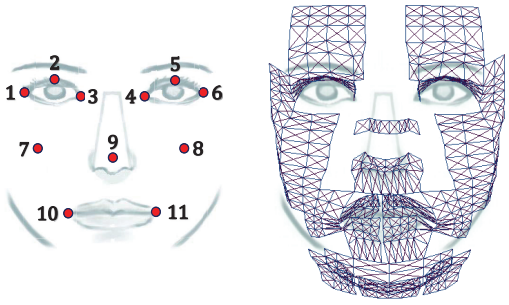
Recent work on static image-based age estimation has considered large-scale evaluations, with 175k images [15]. Obviously, there is still room for improvement, and the area is actively researched. Nonetheless, there is a shortcoming in the literature on the evaluated features for age estimation: Facial expression dynamics are not used at all. This may be due to the lack of proper databases to explore the contribution of dynamic features for age estimation. In this paper, we seek to remedy this shortcoming, and focus on the inclusion of expression dynamics to improve age estimation. To the best of our knowledge, this is the first attempt to investigate and integrate the dynamics of facial features (such as speed, acceleration, amplitude, etc.) for the task of age estimation.

3. METHOD

In this section, details of the proposed age estimation system will be given. The proposed system combines the appearance features with the facial expression dynamics. Our method assumes that the input video starts with a moderately frontal face, and has the entire duration of a smile expression. The flow of the system can be summarized as follows. Initially, 11 facial fiducial points are located in the first frame, and tracked during the rest of the video. Then, the tracked points are used to calculate displacement signals of eyelids, cheeks, and lip corners. Temporal phases (onset, apex, and offset) of the smile are estimated using the mean displacement signal of the lip corners. Afterwards, dynamic features for the eyelid, cheek, and lip corner movements are extracted from each phase. Local Binary Patterns (LBP) features are extracted using the first frame of onset phase, where the face is neutral. After a feature selection procedure, the most informative dynamic features are selected and fused with LBP features to train Support Vector Machine (SVM) classifiers/regressors.

3.1 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)). Each point is initialized in the first frame of the videos for precise tracking and analysis. To track the facial features and pose, we use a piecewise Bézier volume deformation (PBVD) tracker, originally proposed by Tao and Huang [27].



(a) (b)

Figure 1: (a) Used facial feature points with their indices and (b) the 3D mesh model

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. For comparison reasons, we report results for manual and automatic initialization of these features. 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 volumeis 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) can be 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. For more robust tracking, image templates of both the previous frame and the first frame of the sequence are used for matching. 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 projective motion of several points.

3.2 Features

In the proposed system we extract appearance-based and dynamic features from smile videos. In general, a smile can be identified as the upward movement of the mouth corners, which corresponds to Action Unit 12 (AU12) in the facial action coding system (FACS) [6]. In terms of anatomy, the *zygomatic major* muscle contracts and raises the corners of the lips during a smile [7]. In terms of dynamics, smiles 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 (may also be very short) of the expression between onset and offset. Likewise, offset is the final phase from expressive to neutral state. According to Ekman there are dozens of smiles which are different in terms of their appearance and meaning. Ekman also 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 [5].

3.2.1 Extraction of Dynamic Features

To analyze and describe the dynamics of a smile, we extract a set of dynamic features from three different face regions such as eyes, cheeks, and mouth [3]. First of all, 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-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 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:

$$\theta = \arccos \frac{U.\mathcal{N}_\mathcal{P}}{\|U\|\,\|\mathcal{N}_\mathcal{P}\|}, \text{ where } \mathcal{N} = \overrightarrow{\ell_9 c_2} \times \overrightarrow{\ell_9 c_1}. \quad (3)$$

In Equation 3, $\overrightarrow{\ell_9 c_2}$ and $\overrightarrow{\ell_9 c_1}$ 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. However, the estimated pitch (θ_x) angle is a subject-dependent measure, 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) can be 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 as follows:

$$\ell'_i = \left[\ell_i - \frac{c_1+c_2}{2}\right] R_x(-\theta'_x) R_y(-\theta_y) R_z(-\theta_z) \frac{100}{\rho(c1,c2)}, \quad (4)$$

where ℓ'_i is the aligned point and R_x, R_y, and R_z denote the 3D rotation matrices for the given angles. $\rho()$ 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. Since the normalized face is approximately frontal with respect to the camera, we ignore the depth (Z) values of the normalized feature points ℓ'_i, and denote them as l_i.

After the normalization, onset, apex, and offset phases of the smile are detected using the approach proposed by Schmidt *et al.* [24], by calculating the amplitude of the smile as the distance of the right lip corner to the lip center during the smile. Since the faces are normalized, lip center is calculated only once in the first frame. As indicated in [24], the detected smile phases may not necessarily represent the exact definition of smile phases which are defined in [6]. Differently from [24], we estimate the smile 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. It is estimated as:

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

where l^t_i denotes the 2D location of the i^{th} point in frame t. This estimate is smoothed by 4253H-twice method [29]. 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 eyelids and the cheeks, additional amplitude signals are computed. We estimate the (normalized) eyelid aperture $\mathcal{D}_{\text{eyelid}}$ and cheek displacement $\mathcal{D}_{\text{cheek}}$ as follows:

$$\mathcal{D}_{\text{eyelid}}(t) = \frac{\kappa(\frac{l^t_1+l^t_3}{2}, l^t_2)\rho(\frac{l^t_1+l^t_3}{2}, l^t_2) + \kappa(\frac{l^t_4+l^t_6}{2}, l^t_5)\rho(\frac{l^t_4+l^t_6}{2}, l^t_5)}{2\rho(l^t_1, l^t_3)}, \quad (6)$$

$$\mathcal{D}_{\text{cheek}}(t) = \frac{\rho(\frac{l^1_7+l^1_8}{2}, l^t_7) + \rho(\frac{l^1_7+l^1_8}{2}, l^t_8)}{2\rho(l^1_7, l^1_8)}, \quad (7)$$

where $\kappa(l_i, l_j)$ 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. The distance between the eye center and the point on upper eyelid is calculated for both left and right eyes and the estimated values are divided by the length of the eyes. Afterwards, $\mathcal{D}_{\text{eyelid}}$ is calculated as the mean aperture signal of left and right eyes. $\mathcal{D}_{\text{cheek}}$ is defined as the sequence of the mean distances of left and right cheek points to the middle point between two cheeks. Middle point between cheek landmarks is estimated for once in the first frame (neutral face). $\mathcal{D}_{\text{lip}}$, $\mathcal{D}_{\text{eyelid}}$, and $\mathcal{D}_{\text{cheek}}$ are hereafter referred to as amplitude signals. In addition to the amplitudes, speed $\mathcal{V}$ and acceleration $\mathcal{A}$ signals are extracted by computing the first and the second derivatives of the amplitudes, respectively:

$$\mathcal{V}(t) = \frac{d\mathcal{D}}{dt}, \quad (8)$$

$$\mathcal{A}(t) = \frac{d^2\mathcal{D}}{dt^2} = \frac{d\mathcal{V}}{dt}. \quad (9)$$

All the calculated amplitude signals are smoothed by 4253H-twice method [29], and then split into three phases as onset, apex, and offset, which have been 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 1. 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 1, 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. For each face region, three 24-dimensional feature vectors are generated by concatenating these features.

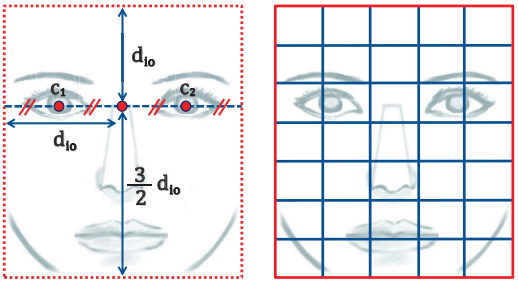
Table 1: Definitions of the extracted features

Feature	Definition
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:	$\frac{(\sum \mathcal{D}^+ - \sum \lvert\mathcal{D}^-\rvert)\omega}{\eta(\mathcal{D})}$

In some cases, 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, the decreasing segments will 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.

3.2.2 *Extraction of Appearance Features*

To describe the appearance of faces, we use uniform *Local Binary Patterns* (LBP) on gray scale images. The original LBP operator, which is proposed by Ojala *et al.* [16], 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 [17], Ojala *et al.* show that the vast majority of the *Local Binary Patterns* in a local neighborhood 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.



(a) (b)

Figure 2: (a) Scaling/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 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, each face is divided into 7×5 non-overlapping (equally-sized) blocks and uniform LBP descriptors are computed on each block (see Fig. 2(b)). 8 neighborhood pixels (on a circle with a radius of 1 pixel) are used to extract the uniform LBP features. All these features are concatenated to form the appearance feature vector. Resultant appearance feature vector is $7 \times 5 \times 59 = 2065$ dimensional. The dimensionality of the appearance feature vectors is reduced by Principal Component Analysis (PCA) so as to retain 99.99% of the variance.

3.3 Feature Selection and Classification

In our system, we use a two-level classification scheme for age estimation, as shown in Fig. 3. In the first level, one-vs-all Support Vector Machine (SVM) classifiers are used to classify the age of a subject into 7 age groups of 10 years ($8-17, 18-27, \ldots, 68-77$). Then, the age of the subject is fine-tuned using an SVM regressor which is specifically trained for the related age group. For a better 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 with the age range (if the estimated age is less/more than the group boundaries, it is set to minimum/maximum age of the group). The resulting estimation of the age is given as an integer with 1 year resolution.

As described in Section 3.2.1, we extract three 24-dimensional 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 [20]. mRMR is an incremental method for minimizing the redundancy while selecting the most relevant information as follows:

$$\max_{f_j \in F - S_{m-1}} \left[I\left(f_j, c\right) - \frac{1}{m-1} \sum_{f_i \in S_{m-1}} I\left(f_j, f_i\right) \right], \quad (10)$$

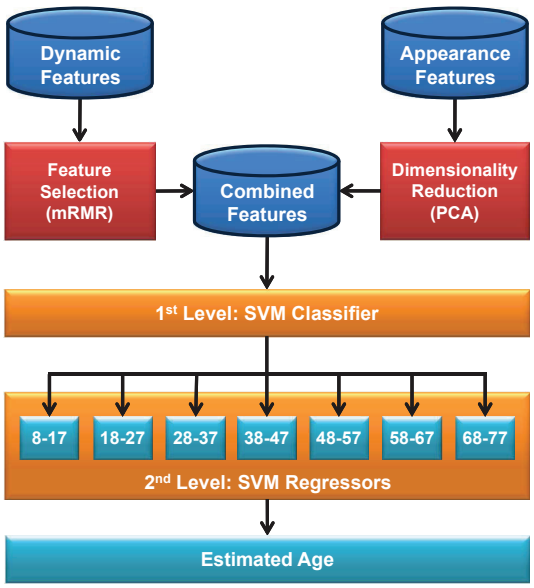


Figure 3: Two-level age estimation architecture using both appearance and dynamic features

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 smile 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, in order to optimize the SVM configuration, different kernels (linear, polynomial, and radial basis function (RBF)) with different parameters (size of RBF kernel, 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.

4. EXPERIMENTAL SETTINGS

4.1 UvA-NEMO Smile Database

The UvA-NEMO Smile Database[1] has been recently collected to analyze the change in dynamics of smiles for different ages [3]. Data collection was carried out in science center NEMO (Amsterdam) as a part of Science Live, the innovative research programme of science center NEMO [25]. NEMO visitors were 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 were 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 further 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). 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.

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 were used to elicit spontaneous smiles. Approximately five minutes of recordings were made per subject, and genuine smiles were segmented.

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

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

4.2 Settings

To evaluate our system and assess the reliability of facial expression dynamics and facial appearance information for age estimation problem, we use the above described smile database of 400 subjects. In our experiments, the two-level classification/regression system is used, and described in section 3.3. The optimum number of selected dynamic features, kernel and parameters of 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. There is no subject overlap between folds. In our experiments, polynomial SVM is found to perform better than linear and RBF alternatives. We initialize the tracking by manually annotated facial landmarks to assess the discrimination power and reliability of individual appearance and dynamic features, and their combination. Additionally, results with automatic initialization are also given for the comparison with other systems, as well as further analysis on different aspects of the age estimation problem.

For automatic facial landmark detection, we use the state-of-the-art system proposed by Dibeklioğlu *et al.* [4]. 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; 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, cheek centers, 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. Correlation coefficients between the extracted amplitude signals with manual and automatic initializations ranged between 0.93 and 1.

5. RESULTS

In this section, we discuss the results of our experiments. First, we will discuss the accuracy of the system when only face dynamics are used, either individually or taken together. Then, we compare these results with the combined use of appearance and dynamics. Finally, we check the effect of gender and expression spontaneity on the accuracy of the system using the combined features.

5.1 Dynamics

Since the proposed dynamic features are extracted from the movements of lip corners, cheeks, and eyelids, we analyze the individual discrimination power of these movements 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. For the reliability of

Figure 4: Sample frames from the UvA-NEMO Smile Database: Showing (a) neutral face, (b) posed enjoyment smile, (c) spontaneous enjoyment smile

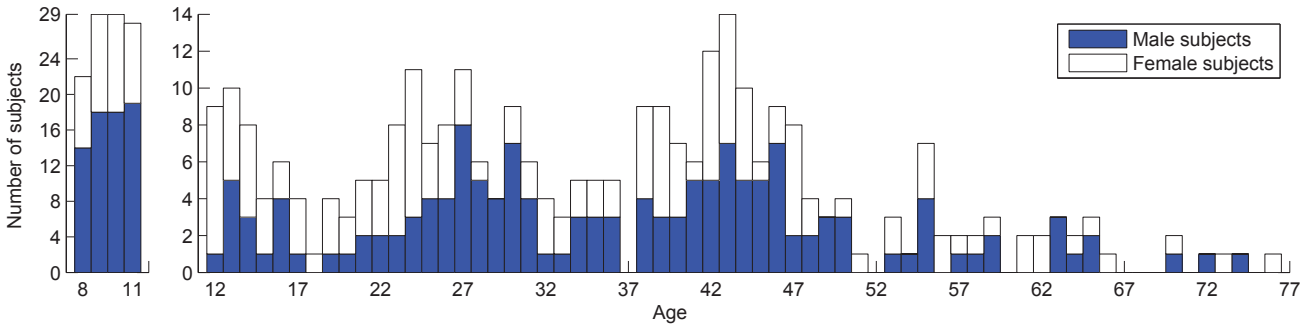


Figure 5: Age and gender distributions of the subjects in the database

results, we use manually initialized tracking in these tests. The resulting *mean absolute error* (MAE) is given in Fig. 6.

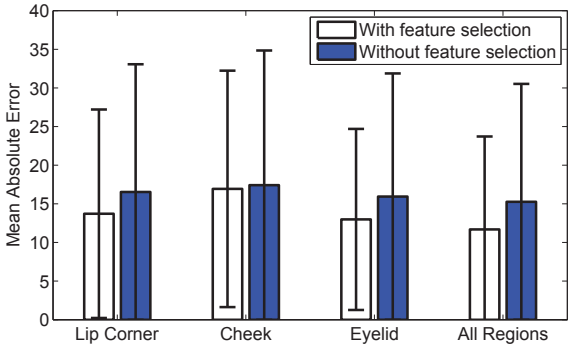


Figure 6: Effect of feature selection on age estimation errors for different facial regions

Our results show that the feature selection increases the accuracy approximately 15% (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 shown that the dynamics of eyelid movements are the most reliable features, with an MAE of 12.98 (± 11.71) years. Lip corner and cheek movement dynamics follow with an MAE of 13.72 (± 13.49) and 16.94 (± 15.30) years, respectively. By combining the dynamic features of different facial regions, the MAE of the age estimation system can be decreased to 11.69 (± 12.02) years.

5.2 Dynamics versus Appearance

In this paper, the aim is to combine facial appearance with expression dynamics for age estimation. However, it is also important to show the discriminative power of facial expression dynamics and appearance, individually. For this purpose, we evaluate the individual and combined use of these features. To assess the effect of tracking initialization on accuracy, we use both manually and automatically annotated facial landmarks to initialize facial tracking in our experiments.

As shown in Table 2, using automatic initialization for tracking increases the MAE by 4.28% and 1.26% for dynamic and appearance features, respectively. Since the reliability of dynamic features is directly related to the accuracy of tracking, automatic initialization affects it more than appearance features. The MAE for combined features is also increased by 2.34% by automatic initial-

Table 2: Mean Absolute Errors for dynamic, appearance, and combined features

Features	Mean Absolute Error	
	Manual Initialization	Automatic Initialization
Dynamics	11.21 ($\pm$11.34)	11.69 ($\pm$12.02)
Appearance	5.57 ($\pm$5.86)	5.64 ($\pm$5.90)
Combination	4.71 ($\pm$4.75)	4.82 ($\pm$4.87)

ization of tracking. However, the use of combined features significantly ($p < 0.001$, Cohen's $d > 0.15$) improve the age estimation accuracy in comparison to the individual use of dynamic and appearance features. Therefore, automatically initialized tracking is used in the remainder of our experiments.

When we analyze the results, it is clear that using only facial dynamics is not enough for an accurate age estimation system. The MAE of using dynamic features is 11.69 ($\pm$12.02) years, where the MAE for facial appearance is only 5.64 ($\pm$5.90) years. Nevertheless, by combining the dynamic and appearance features, the proposed system is able to achieve the best result with an MAE of 4.82 ($\pm$4.87), which is significantly ($p < 0.001$, Cohen's $d > 0.15$) more accurate than using dynamic and appearance features individually.

5.3 Effect of Gender

To assess the effect of gender on the accuracy of the system using the combined features (with automatic initialization of tracking), a gender-specific age estimation system is implemented and compared with the general method. In the gender-specific system, 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 correctly given. The MAEs for both gender-specific and general training are given in Table 3.

Table 3: Comparison of the gender-specific method with the general method for age estimation

Method	Mean Absolute Error		
	Male	Female	Mean
Specific	4.26 ($\pm$3.36)	5.01 ($\pm$5.73)	4.63 ($\pm$4.70)
General	4.59 ($\pm$4.80)	5.03 ($\pm$4.95)	4.81 ($\pm$4.87)

Our results show that the gender-specific training decreases the overall MAE of the system from 4.81 ($\pm$4.87) years to 4.63 ($\pm$4.70) years. In particular, the gender-specific training decreases the MAE by 7.19% for males, and 0.40% for females, with respect to training on the whole data. These results show the difference of the proposed combined features between males and females. Fig. 7 shows the estimated ages using the general and the gender-specific methods.

5.4 Effect of Expression Spontaneity

In order to assess the effect of expression spontaneity on the accuracy of using combined features, a spontaneity-specific age estimation system is also constructed. Here we compare its accuracy with the general approach (using automatic initialization of tracking). For this purpose, separate classifiers/regressors are trained for spontaneous and posed smiles. It is assumed that the spontaneity of all smiles are correctly classified and the same training procedure is separately applied to each subset.

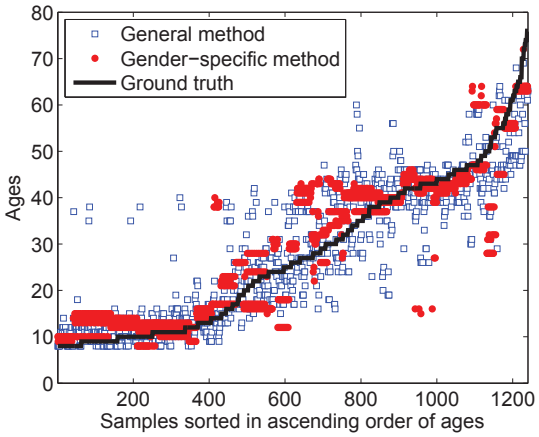


Figure 7: Estimated ages using the general and the gender-specific methods

Table 4: Comparison of the spontaneity-specific method with the general method for age estimation

Method	Mean Absolute Error		
	Spontaneous	Posed	Mean
Specific	4.51 ($\pm$5.61)	4.02 ($\pm$3.67)	4.26 ($\pm$4.71)
General	4.92 ($\pm$5.09)	4.70 ($\pm$4.66)	4.81 ($\pm$4.87)

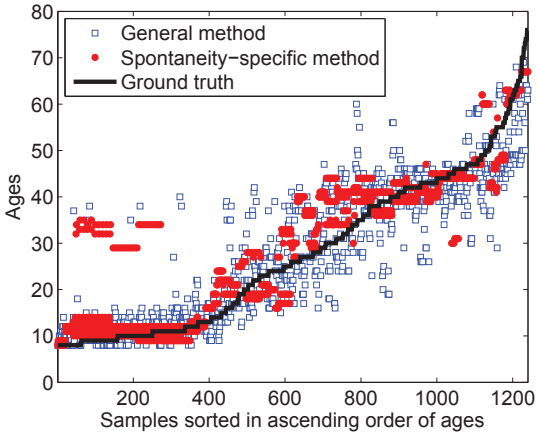


Figure 8: Estimated ages using the general and the spontaneity-specific methods

As shown in Table 4, spontaneity-specific approach performs with an MAE of 4.26 ($\pm$4.71), therefore improving the accuracy by 11.43% with respect to the general approach. 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. Plot of the estimated ages using the general and the spontaneity-specific methods are shown in Fig. 8.

5.5 Comparison with other methods

To the best of our knowledge, this is the first study which uses the facial expression dynamics (such as speed, acceleration, amplitude, etc.) for age estimation problem. Except a recent work [12], none of the previous studies in the literature focus on using temporal information for age estimation.

In [12], Hadid proposes to use spatiotemporal 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). [12] uses 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 each 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 spatiotemporal approach reaches only 69.2%.

VLBP is a straightforward extension of original LBP operator to describe dynamic textures (image sequences) [32]. VLBP enables the use of temporal space (T), looks the face sequence as a volume, and the neighborhood of each pixel is defined in three dimensional space, where LBP uses only X and Y dimensions of a single image. Then, the histograms of VLBP are used as features. In [32], Zhao *et al.* 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 the related approaches, we implement three baseline methods: (1) VBLP-based spatiotemporal approach, (2) spatiotemporal approach using VBLP-TOP features, and (3) appearance-based approach which classifies each frame in a video using LBP features with SVM classifiers, individually, and fuses the classification results by majority voting. Both LBP-TOP- and VBLP-based methods use the same classification/regression architecture with our method. 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 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 VBLP. Time interval for volumetric approaches is set to three frames. Zhao *et al.* [32] have shown that these neighborhood and time interval parameters perform well on facial expression classification. The dimensionality of LBP, VLBP, and LBP-TOP features is reduced by Principal Component Analysis (PCA) so as to retain 99.99% of the variance.

The cumulative error distributions of the mean absolute error for different methods are given in Fig. 9 (a). Given error distributions show that the proposed system, which uses both dynamic and appearance features, significantly outperforms all other methods ($p < 0.001$, Cohen's $d > 0.15$). Additionally, Figs. 9 (b-f) show the comparison of the estimated ages using the combined features and the other methods.

As shown in Table 5, the combination of dynamic and appearance features achieves the minimum MAE of 4.81 (±4.87) years. It is important to indicate that using LBP features extracted on a single, automatically detected neutral image provides more accurate age estimation than using all frames in a video and voting on the results. Spatiotemporal methods can only reach a mean accuracy of 17.03 (±13.56) and 14.95 (±11.41) years with VLBP and LBP-TOP features, respectively. By using the proposed dynamic features only, the system is significantly more accurate than when it uses the spatiotemporal features ($p < 0.001$, Cohen's $d > 0.25$). Note that we assume that an age estimate is correct if the error from the ground truth is less than 10 years. In this case, the proposed system which combines appearance and dynamics features improves the correct classification rate by 4.36% (absolute) with respect to the appearance features only.

6. DISCUSSION

In our experiments, we show that dynamics of the eyelid movements perform best for individual regions. Additionally, fusion of eyelid, lip corner, and cheek movement dynamics (with a feature selection step) improves the accuracy of the eyelid dynamics by 9.94%. For dynamic features, using feature selection increases the accuracy approximately by 15% on average, as well as reducing feature dimensionality. This finding indicates that there is a significant amount of noise or confusing information in dynamic features.

Our results show that the individual use of the facial expression dynamics is not enough for 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 improved ($p < 0.001$, Cohen's $d > 0.15$) by enabling the dynamics of smile expression. Moreover, the proposed combined features outperform all the baseline methods tested in this study. These results confirm the importance of the information hidden in facial expression dynamics.

When we analyze the effect of initialization on facial tracking, it is seen that automatic initialization decreases the reliability of dynamic features approximately three times more than the appearance features, since the reliability of dynamic features are directly related to the accuracy of tracking. However, under both automatic and manual initialization conditions, the use of the combined features significantly ($p < 0.001$, Cohen's $d > 0.15$) improves the age estimation accuracy in comparison to the individual use of dynamic and appearance features.

Using multivariate analysis of variance (MANOVA), we find the most selected dynamic features and significant ($p < 0.001$, $\eta^2 > 0.10$) feature differences between different ages. Our findings indicate that the dynamics of smile onsets are more discriminative than other smile phases for age estimation. 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. Then, we analyzed the significant ($p < 0.001$) differences of these features between spontaneous and posed smiles using the t-test. Our results show that the maximum speed and the acceleration of the lip corner movements are significantly higher for posed smiles during smile onsets ($p < 0.001$, Cohen's $d > 0.40$). Also, the maximum speed and the acceleration of eye closure are significantly higher for spontaneous smiles in smile onsets ($p < 0.001$, Cohen's $d > 0.20$). These findings can explain the higher accuracy of the spontaneity-specific system. Similarly, t-test analysis is repeated for male and female differences. However, no significant difference (in the related dynamic features) is found between male and female subjects. This finding indicates that the

Table 5: Mean Absolute Error for different methods

Method | Age Range	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	Total
Proposed: Dynamics	6.20	8.18	13.55	15.22	10.82	20.62	24.07	29.06	11.69 ($\pm$12.02)
Proposed: Appearance	2.57	3.77	6.53	7.71	5.37	9.91	12.47	13.76	5.64 ($\pm$5.90)
Proposed: Combined	2.73	2.99	5.45	6.83	4.35	8.45	10.87	13.18	4.81 ($\pm$4.87)
Appearance: LBP, Voting	3.27	4.39	7.24	8.10	5.75	11.00	12.83	15.65	6.24 ($\pm$6.44)
Spatiotemporal: VLBP	16.99	15.09	13.28	13.51	16.71	29.83	39.97	53.00	17.03 ($\pm$13.56)
Spatiotemporal: LBP-TOP	15.59	13.48	11.04	11.76	15.16	25.11	33.10	44.94	14.95 ($\pm$11.41)
Number of Samples	158	333	215	171	250	66	30	17	1240

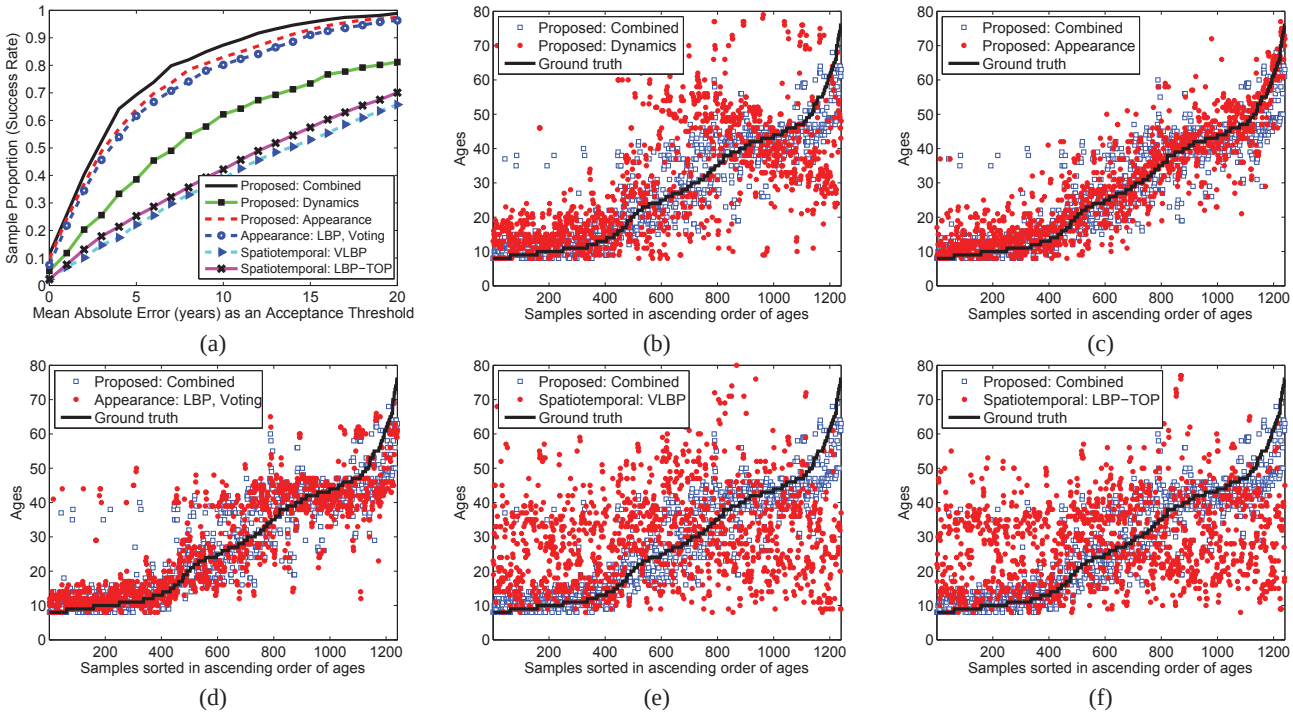


Figure 9: (a) Cumulative error distribution of the mean absolute error for different methods. (b-f) Comparison of the estimated ages using different methods

higher accuracy of gender-specific system is mostly based on LBP-based appearance features.

Experimental results show that spatiotemporal approaches based on VLBP and LBP-TOP are not efficient for age estimation. Even the individual use of our dynamic features outperforms these methods, significantly. Spatiotemporal features describe the change of facial appearance in time, but our proposed method models the appearance on a single neutral image (which is automatically selected as the first frame of the onset phase) and adds the dynamics of the facial expression (such as amplitude, speed, acceleration, etc) on it. As a result, the proposed system (using combined features) is significantly ($p < 0.001$, Cohen's $d > 0.15$) more accurate than all the competitor methods in our experiments.

7. CONCLUSIONS

In this study, we have introduced the usage of dynamic features to improve age estimation. Previously considered methods in the literature evaluate the appearance of the face, as the appearance is the most revealing aspect of aging. However, we have shown that the speed and movement features of facial muscles during smiling can improve the estimation of a person's age. While such dynamic features by themselves are not as accurate as appearance features, their contribution is significant when combined.

We assessed a range of dynamical features in an exploratory fashion, as geriatrics literature on facial aging also mostly focuses on appearance features. Therefore, we were not able to find sufficiently detailed descriptions of age effects on muscle dynamics (except for a study that reports intact muscle dynamics for orbicularis oculi, which contradicts some earlier studies [21]). Among individual regions of the face, our results show that eyelid dynamics are the most revealing in terms of age estimation, followed by lip corners and cheeks. We did not find a significant effect of the proposed features over different genders.

Our results are derived from a recently collected database with 400 subjects aged between 8 to 76 years. This is the most extensive dynamic age evaluation study to this date in the literature. We con-

trast dynamic information obtained from spontaneous and posed smiles. Posed expressions are somewhat better for age estimation, but not significantly so. We also compare our method to an appearance based baseline, as well as to a recently proposed spatiotemporal approach.

8. ACKNOWLEDGMENTS

This research was part of Science Live, the innovative research programme of science center NEMO that enables scientists to carry out real, publishable, peer-reviewed research using NEMO visitors as volunteers. Additionally, this study is partially supported by Boğaziçi University project BAP-6531.

9. 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 Science International*, 172(1):1–9, 2007.

[2] T. R. Alley. *Social and applied aspects of perceiving faces*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[3] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. In *ECCV*, 2012.

[4] H. Dibeklioğlu, A. A. Salah, and T. Gevers. A statistical method for 2-d facial landmarking. *IEEE Trans. on Image Processing*, 21(2):844–858, 2012.

[5] P. Ekman. *Telling lies: Cues to deceit in the marketplace, politics, and marriage*. New York: WW. Norton & Company, 1992.

[6] P. Ekman and W. V. Friesen. *The Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press Inc., San Francisco, CA, 1978.

[7] P. Ekman and W. V. Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6:238–252, 1982.

[8] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Trans. on PAMI*, 32(11):1955–1976, 2010.

[9] X. Geng, K. Smith-Miles, and Z. Zhou. Facial age estimation by nonlinear aging pattern subspace. In *ACM Multimedia*, pages 721–724, 2008.

[10] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *ACM Multimedia*, pages 307–316, 2006.

[11] 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. on Image Processing*, 17(7):1178–1188, 2008.

[12] A. Hadid. Analyzing facial behavioral features from videos. In A. A. Salah and B. Lepri, editors, *Int. Workshop on Human Behavior Understanding*, pages 52–61, 2011.

[13] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999.

[14] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. on PAMI*, 24(4):442–455, 2002.

[15] B. Ni, S. Yan, et al. Web image and video mining towards universal and robust age estimator. *IEEE Trans. on Multimedia*, 13(6):1217–1229, 2011.

[16] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on PAMI*, 24(7):971–987, 2002.

[18] M. Ortega, L. Brodo, M. Bicego, and M. Tistarelli. On the quantitative estimation of short-term aging in human faces. In *ICIAP*, pages 575–584, 2009.

[19] A. J. O'Toole, T. Vetter, H. Volz, and E. Salter. Three-dimensional caricatures of human heads: Distinctiveness and the perception of age. *Perception*, 26:719–732, 1997.

[20] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on PAMI*, 27(8):1226–1238, 2005.

[21] F. Pottier, N. Z. El-Shazly, and A. E. El-Shazly. Aging of orbicularis oculi. *Archives of Facial Plastic Surgery*, 10(5):346–349, 2008.

[22] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144, 2009.

[23] R. Sanders. Torsional elasticity of human skin in vivo. *Pflügers Archiv European Journal of Physiology*, 342(3):255–260, 1973.

[24] K. L. Schmidt, J. F. Cohn, and Y. Tian. Signal characteristics of spontaneous facial expressions: automatic movement in solitary and social smiles. *Biological Psychology*, 65(1):49–66, 2003.

[25] Science Center NEMO. http://www.e-nemo.nl/.

[26] J. Suo, S. C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *IEEE Trans. on PAMI*, 32(3):385–401, 2010.

[27] H. Tao and T. Huang. Explanation-based facial motion tracking using a piecewise Bezier volume deformation model. In *CVPR*, volume 1, pages 611–617, 1999.

[28] B. Tiddeman, M. Burt, and D. I. Perrett. Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, 21(5):42–50, 2001.

[29] P. F. Velleman. Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, pages 609–615, 1980.

[30] Y. Wu, N. Thalmann, and D. Thalmann. A dynamic wrinkle model in facial animation and skin aging. *Journal of Visualization and Computer Animation*, 6:195–205, 1995.

[31] C. Zhan, W. Li, and P. Ogunbona. Age estimation based on extended non-negative matrix factorization. In *IEEE Int. Workshop on Multimedia Signal Processing*, 2011.

[32] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. on PAMI*, 29(6):915–928, 2007.