Gender estimation based on smile-dynamics

Antitza Dantcheva, and François Brémond

Abstract—Automated gender estimation has numerous applications including video surveillance, human computer-interaction, anonymous customized advertisement and image retrieval. Most commonly, the underlying algorithms analyze the facial appearance for clues of gender. In this work we propose a novel method for gender estimation, namely the use of dynamic features gleaned from smiles and proceed to show that (a) facial dynamics can be used to improve appearance-based gender estimation, (b) that while for adult individuals appearance features are more accurate than dynamic features, for subjects under 18 years old facial dynamics can outperform appearance features. In addition, we fuse proposed dynamics-based approach with state-of-the-art appearance based algorithms, predominantly improving appearance-based gender estimation performance. Results show that smile-dynamics include pertinent and complementary to appearance gender information.

Keywords—soft biometrics, gender estimation, facial dynamics.

I. INTRODUCTION

Human facial analysis has engaged researchers in multiple fields including computer vision, biometrics, forensics, cognitive psychology and medicine. Interest in this topic has been fueled by scientific advances that provide insight into a persons identity, intent, attitude, aesthetics as well as health, solely based on their face images.

Besides establishing an individuals identity, ancillary information may also be gleaned from face images related to personal attributes such as gender, age and ethnicity. Gender and specifically automated gender estimation has been of specific interest for its broad application range, be it in surveillance [67], human computer-interaction, anonymous customized advertisement systems1 or image retrieval systems [6], leading to numerous commercial applications234. Also, gender has been a prominent soft-biometric trait [21], [23], which can be employed (a) in fusion with other biometric traits to improve the matching accuracy of a biometric system [45], (b) in fusion with other soft biometrics for person authentication [19], [20], or (c) as a filter for search space reduction [22].

Automated gender estimation remains a challenging research area, due to large intra-class variation [52], and also due to challenges concerning illumination, as well as pose, age and ethnicity of a person. Further, facial expressions have a negative effect on the accuracy of automated gender estimation systems. This is why the majority of previous works have extracted and studied appearance-based features under the simplifying assumption of neutral face expressions with reasonably good results.

A. Gender and emotional expression

Deviating from such works, we here introduce the usage of a set of dynamic facial features for gender estimation. Specifically, we focus on extracting dynamic features from a common facial expression, namely the smile, and study how smile-dynamics encrypt gender evidence. The hypothesis is that male and female smile-dynamics differ in parameters such as intensity and duration. This hypothesis is supported in part by a number of cognitive-psychological studies, showing evidence for gender-dimorphism in the human expression [14], [74], [41], [1], [51], [25]. A main observation of such studies has been that females express emotions more frequently than males, and in the context of smile, females tend to smile more often than men in a variety of social contexts [25]. Such observations follow the theorem of men exhibiting restrictive emotionality and thus being unwilling to self-disclose intimate feelings. It is interesting to note, that a gender-based difference in emotional expression is observed as early as in 3 months old, shaped by how caregivers interact to male and female infants [33]; and also observed in toddlers, which appears to be further trained in social interactions [16], [56], [57]. Moreover, females are more accurate expressers of emotion, when posing deliberately and when observed unobtrusively, which is consistent across cultures [11]. The same work assigns happiness and fear as male.gender-stereotypical expressions. On the other hand, faces showing anger are considered more masculine [41], [40], [42], [43], [4], [5], [83] in the context of human gender recognition.

B. Contributions

Motivated from the above, we propose the use of an automated framework for facial dynamics extraction based on signal displacement of facial distances between key facial landmarks. We analyze the properties of 27 such facial distances in smile-video-sequences with emphasis on spontaneous, as well as posed smiles. The proposed dynamic features are fully complementary to appearance based features, and when combined with appearance, can pose an increased difficulty for spoof-attacks. We have adopted the approach from Dibeklioğlu et al. [27], [28], where it has been used for age estimation, as well as spontaneous vs. posed smile detection based on facial dynamics, see also [29], [26].

The use of the framework is instrumental in answering following questions:

- Do facial dynamics provide information about gender in (a) spontaneous smile- and (b) posed smile video sequences?
• Can facial smile dynamics improve the accuracy of appearance based gender estimation systems?
• Which gender can pose smiles more genuinely?

Related work of a holistic smile-based gender estimation algorithm can be found in Biliński et al. [9].

C. Structure of paper

This work is organized as follows: Section I-D revisits existing works on gender estimation. Section II proceeds to describe the proposed method, elaborating on individual steps (face detection, landmark location, selected features, statistics of dynamic features, feature selection, classification and used appearance features). Section III presents the employed dataset and the subsequent Section IV depicts and discusses related experimental results. Finally Section V concludes the paper.

D. Related work

Gender estimation Existing introductory overviews for algorithms related to gender estimation include the works of Ng et al. [61], Bekios-Calfa et al. [7], Ramanathan et al. [65], Mäkinen and Raisamo [54] and Dantcheva et al. [21]. Based on these works we can conclude that gender estimation remains a challenging task, which is inherently associated with different biometric modalities including fingerprint, face, iris, voice, body shape, gait, signature, DNA, as well as clothing, hair, jewelry and even body temperature. The forensic literature [52] suggests that the skull, and specifically the chin and the jawbone, as well as the pelvis, are the most significant indicators of the gender of a person; in juveniles, these shape-based features have been recorded to provide classification accuracy of 91% – 99%.

Humans are generally quite good at gender recognition from early in life (e.g., [62], [64]), probably reflecting evolutionary adaptation. As pointed out by Edelman et al. [30], humans perform facial image based gender classification with an error rate of about 11%, which is commensurate to that of a neural network algorithm performing the same task.

Dynamics have been used in the context of body-based classification of gender. Related cues include body sway, waist-hip ratio, and shoulder-hip ratio (see [59]); for example, females have a distinct waist-to-hip ratio and swing their hips more, whereas males have broader shoulders and swing their shoulders more.

Despite these recent successes, automated gender recognition from biometric data remains a challenge and is impacted by other soft biometrics, for example, age and ethnicity; gender dimorphism is accentuated only in adults, and varies across different ethnicities.

Automated Image-based Gender Estimation from Face

In gender estimation from face, feature-based approaches extract and analyze a specific set of discriminative facial features (patches) in order to identify the gender of a person. This is a particularly challenging problem, as is implied from the fact that female and male average facial shapes are generally found to be very similar [50].

Another challenge comes to the fore in unconstrained settings with different covariates, such as illumination, expressions and ethnicity. While in more constrained settings, face-based gender estimation has been reported to achieve classification rates of up to 99.3% (see Table I), this performance though significantly decreases in more realistic and unconstrained settings.

The majority of gender classification methods contain two steps preceding face detection, namely feature extraction and pattern classification.

Feature extraction: Notable efforts include the use of SIFT [75], LBP [54], semi-supervised discriminant analysis (SDA) [8] or combinations of different features [36], [79].

Classification: A number of classification methods have been used for gender estimation, and a useful comparative guide of these classification methods can be found in Mäkinen and Raisamo [55]. One interesting conclusion of their work was that image size did not greatly influence the classification rates. This same work also revealed that manual alignment affected the classification rates positively, and that the best classification rates were achieved by SVM.

The area of gender estimation has also received some other contributions such as those that go beyond using static 2D visible spectrum face-images. Interesting related work include the work of Han et al. [39], exploring 3D images, Gonzalez–Sosa et al. [35], studying jointly body and face, and Chen and Ross [18], [69], using near-infrared (NIR) and thermal images for gender classification.

Expression Recognition Automated expression recognition has received increased attention in the past decade, since it is particularly useful in a variety of applications, such as human computer interaction, surveillance and crowd analytics. The majority of methods aim to classify 7 universal expressions namely neutral, happy, surprised, fearful, angry, sad, and disgusted [82] based on the extracted features used. Classi-cal approaches follow Ekman’s facial action coding system (FACS) [31], assigning each facial unit to represent movement of a specific facial muscle. In this context, intensity and number of facial units have been studied, as well as of action unit combinations, towards expression recognition. Interesting work can be found in related survey papers [84], [58], [71] and in a related recent expression-recognition-challenge-study [76]. Latest advances involve deep learning [85], [47].

Inspired by cognitive, psychological and neuroscientific findings, facial dynamics have been used previously towards improving face recognition [38], gender estimation [24], age estimation [27], as well as kinship recognition reported in a review article by Hadid et al. [37].

II. Dynamic Feature Extraction in Smile-Video-Sequences

Deviating from the above works on gender estimation, we propose to extract dynamic features in smile-video-sequences. The general scheme is shown in Fig. 1. Specifically we focus on signal displacement of facial landmarks, as we aim to study among others the pertinence of different facial landmarks, as well as the pertinence of different statistical properties of facial
TABLE I. OVERVIEW OF FACE-BASED GENDER CLASSIFICATION ALGORITHMS. ABBREVIATIONS USED: PRINCIPAL COMPONENT ANALYSIS (PCA), INDEPENDENT COMPONENT ANALYSIS (ICA), SUPPORT VECTOR MACHINES (SVM), GAUSSIAN PROCESS CLASSIFIERS (GPC), ACTIVE APPEARANCE MODEL (AAM), LOCAL BINARY PATTERN (LBP), ACTIVE SHAPE MODEL (ASM), DISCRETE COSINE TRANSFORM (DCT), SEMI-SUPERVISED DISCRIMINANT ANALYSIS (SDA).

<table>
<thead>
<tr>
<th>Work</th>
<th>Features</th>
<th>Classifier</th>
<th>Datasets used for evaluation</th>
<th>Performance numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bekios-Calfa et al. (2007) [7]</td>
<td>PCA</td>
<td>SVM</td>
<td>UCN (nonpublic), 10,700 images</td>
<td>93.46% ± 1.65%</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td></td>
<td>FERET, 994 images</td>
<td>93.57% ± 1.39%</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td></td>
<td>PAL, 576 images</td>
<td>93.57% ± 1.39%</td>
</tr>
<tr>
<td>Xia et al. (2008) [80]</td>
<td>LBP, Gabor</td>
<td>SVM</td>
<td>CAS-PEAL, 10,784 images</td>
<td>93.74%</td>
</tr>
<tr>
<td>Mäkinen and Raisamo (2008) [55]</td>
<td>LBP</td>
<td>SVM</td>
<td>FERET, 411 images</td>
<td>86.54%</td>
</tr>
<tr>
<td>Gao and Ai (2009) [34]</td>
<td>ASM</td>
<td>Adaboost</td>
<td>Private, 1,300 images</td>
<td>92.89%</td>
</tr>
<tr>
<td>Toews and Arbel (2009) [75]</td>
<td>SIFT</td>
<td>Bayesian</td>
<td>FERET, 994 images</td>
<td>83.7%</td>
</tr>
<tr>
<td>Shan (2010) [72]</td>
<td>LBP</td>
<td>Adaboost</td>
<td>LFW, 7,443 images</td>
<td>94.44%</td>
</tr>
<tr>
<td>Guo et al. (2009) [36]</td>
<td>LBP, HOG, BIF</td>
<td>SVM</td>
<td>YGA, 8,000 images</td>
<td>89.28%</td>
</tr>
<tr>
<td>Wang et al. (2010) [79]</td>
<td>SIFT, context</td>
<td>Adaboost</td>
<td>FERET, 2,409 images</td>
<td>95.0%</td>
</tr>
<tr>
<td>Nazhir et al. (2010) [60]</td>
<td>DCT</td>
<td>KNN</td>
<td>SUMS, 400 images</td>
<td>99.3%</td>
</tr>
<tr>
<td>Ross and Chen (2011) [69]</td>
<td>LBP</td>
<td>SVM</td>
<td>CBSR NIR, 3,200 images</td>
<td>93.59%</td>
</tr>
<tr>
<td>Cao et al. (2011) [12]</td>
<td>Metrology</td>
<td>SVM</td>
<td>MUCT, 276 images</td>
<td>86.83%</td>
</tr>
<tr>
<td>Hu et al. (2011) [44]</td>
<td>Filter banks</td>
<td>SVM</td>
<td>Flickr, 26,700 images</td>
<td>90.1%</td>
</tr>
<tr>
<td>Bekios-Calfa et al. (2011) [8]</td>
<td>SDA</td>
<td>PCA</td>
<td>Multi-PIE, 337 images</td>
<td>88.04%</td>
</tr>
<tr>
<td>Shan (2012) [73]</td>
<td>Boosted LBP</td>
<td>SVM</td>
<td>LFW, 7,443</td>
<td>94.81%</td>
</tr>
<tr>
<td>Ramón-Balmaseda (2012) [66]</td>
<td>LBP</td>
<td>SVM</td>
<td>MORPH, LFW, Images of Groups, 17,814</td>
<td>75.10%</td>
</tr>
<tr>
<td>Jia and Cristianini (2015) [46]</td>
<td>Multi-scale LBP</td>
<td>C-Pegasos</td>
<td>Private, 4 million images</td>
<td>96.86%</td>
</tr>
</tbody>
</table>

Fig. 1. Proposed framework for automatic gender estimation.

dynamics (e.g. intensity and duration) in the effort of gender estimation.

Towards extraction of such dynamic features, we assume a near frontal pose of the subject and an initial near-neutral expression of the subject (given in the used dataset).

A. Face Detection and Extraction of Facial Landmarks

Firstly we detect the face using the well established Viola and Jones algorithm [78]. We here note that the faces were robustly detected in all video sequences and frames. Within the detected face we identify facial feature points corresponding to points in the regions of the eye brows, eyes, nose and lips (see Fig. 5). Specifically we employ the facial landmark detection algorithm proposed in the work of Asthana et al. [2]. The algorithm is an incremental formulation for the discriminative deformable face alignment framework [81], using a discriminative 3D facial deformable shape model fitted to a 2D image by a cascade of linear regressors. The detector was trained on the 300W dataset (a dataset introduced in the context of the 300 faces in-the-wild challenge [70]) and detects 49 facial landmarks (see Fig. 5). For the UvA Nemo-dataset the facial landmarks were detected robustly in all video sequences and frames. We use these points to initialize a sparse optical flow tracking algorithm, based on the Kanade-Lucas-Tomasi (KLT) algorithm [53] in the first frame of each video-sequence. For the here proposed framework we select a subset of facial-points in three different face regions: (a) eye brow region, (b) eye region, (c) mouth region (see Fig. 2) and proceed to extract dynamic features thereof.

B. Extraction of Dynamic Features

We extract dynamic features corresponding to the signal-displacement in facial-distances depicted in Table II. We have selected 27 such facial-distances based on findings on facial
movements during smile-expressions [68].

1) Temporal smile-segmentation: Generally, the human smile is caused by the contraction of the zygomatic major muscle, which raises the corners of the lips [32], corresponding to “Action Unit Nr. 12” in Ekman’s facial action coding system [31]. Temporally segmented, the human smile contains three phases: (a) onset: contraction of the zygomatic major muscle and alteration from neutral to expressive state, (b) apex: peak period of the expressive state, and (c) offset: relaxation of the zygomatic major muscle and change from expressive to neutral state. We here note that there are dozens of smile-classes, differing in appearance and meaning.

The next step in our method is to temporally segment the signal-displacement functions as: (a) onset: duration of monotonous increase, (b) apex: phase between onset and offset, (c) offset: duration of monotonous decrease. Fig. 3 illustrates two examples of signal-displacement in the mouth-region ($D_2$, mouth length), leading to a smile-curve with differently pronounced onset, apex and offset phases.

We smoothen each of the 27 signal displacement functions by the 4253H-twice smoothing algorithm [77] to flatten minor tracking-flaws.

2) Statistics of Dynamic Features: We proceed to extract statistics from each dynamic function with respect to the particular smile-phases, denoted by the superindices ($^*$) for onset, ($^a$) for apex, and ($^o$) for offset, which we summarize in Table III. We compute the speed as $V(t) = \frac{dD}{dt}$ and the acceleration as $A = \frac{d^2D}{dt^2} = \frac{dV}{dt}$. We denote the number of frames by $\eta$, frame rate of the video sequence by $\omega$. Each of the defined 27 signal-displacement-functions are represented by a set of 24 features, resulting in a 648-dimensional feature vector.

C. Feature Selection

We use the Min-Redundancy Max-Relevance (mRMR) algorithm [63] for selecting the permanent dynamic proposed features. mRMR minimizes the redundancy, while selecting

<table>
<thead>
<tr>
<th>Facial Distance</th>
<th>Description</th>
<th>Description by facial landmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>Width of right eye</td>
<td>$\rho(\frac{l_{14}+l_{21}+l_{17}+l_{18}}{l_{11}+l_{12}})$</td>
</tr>
<tr>
<td>$D_2$</td>
<td>Width of left eye</td>
<td>$\rho(\frac{l_{13}+l_{16}}{l_{11}+l_{12}})$</td>
</tr>
<tr>
<td>$D_3$</td>
<td>Length of right eye</td>
<td>$\rho(l_{11}, l_{10})$</td>
</tr>
<tr>
<td>$D_4$</td>
<td>Length of left eye</td>
<td>$\rho(l_{17}, l_{10})$</td>
</tr>
<tr>
<td>$D_5$</td>
<td>Length of mouth</td>
<td>$\rho(l_{15}, l_{24})$</td>
</tr>
<tr>
<td>$D_6$</td>
<td>Width of mouth</td>
<td>$\rho(l_{22}, l_{25})$</td>
</tr>
<tr>
<td>$D_7$</td>
<td>Center of mouth to left side of upper lip</td>
<td>$\rho(\frac{l_{22}+l_{24}+l_{26}}{l_{21}})$</td>
</tr>
<tr>
<td>$D_8$</td>
<td>Center of mouth to right side of upper lip</td>
<td>$\rho(\frac{l_{22}+l_{24}+l_{23}}{l_{24}})$</td>
</tr>
<tr>
<td>$D_9$</td>
<td>Center of mouth to left mouth corner</td>
<td>$\rho(\frac{l_{22}+l_{24}}{l_{19}})$</td>
</tr>
<tr>
<td>$D_{10}$</td>
<td>Center of mouth to right mouth corner</td>
<td>$\rho(\frac{l_{22}+l_{24}}{l_{27}})$</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>Center of mouth to upper lip</td>
<td>$\rho(\frac{l_{22}+l_{24}}{l_{10}+l_{24}})$</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>Center of mouth to average distance of two mouth corners</td>
<td>$\rho(\frac{l_{14}+l_{20}}{l_{11}})$</td>
</tr>
<tr>
<td>$D_{13}$</td>
<td>Left side of right eyebrow to nose</td>
<td>$\rho(l_{13}, l_{26})$</td>
</tr>
<tr>
<td>$D_{14}$</td>
<td>Right side of left eyebrow to nose</td>
<td>$\rho(l_{13}, l_{26})$</td>
</tr>
<tr>
<td>$D_{15}$</td>
<td>Center of right eyebrow to right side of the right eyebrow</td>
<td>$\rho(l_{15}, l_{46})$</td>
</tr>
<tr>
<td>$D_{16}$</td>
<td>Center of right eyebrow to left side of the right eyebrow</td>
<td>$\rho(l_{15}, l_{41})$</td>
</tr>
<tr>
<td>$D_{17}$</td>
<td>Center of left eyebrow to right side of the left eyebrow</td>
<td>$\rho(l_{21}, l_{45})$</td>
</tr>
<tr>
<td>$D_{18}$</td>
<td>Center of left eyebrow to left side of the left eyebrow</td>
<td>$\rho(l_{21}, l_{41})$</td>
</tr>
<tr>
<td>$D_{19}$</td>
<td>Distance between eyebrows</td>
<td>$\rho(l_{13}, l_{41})$</td>
</tr>
<tr>
<td>$D_{20}$</td>
<td>Left corner of left eye to left mouth corner</td>
<td>$\rho(l_{17}, l_{49})$</td>
</tr>
<tr>
<td>$D_{21}$</td>
<td>Right corner of left eye to center of mouth</td>
<td>$\rho(l_{10}, l_{22}+l_{24})$</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>Left corner of right eye to center of mouth</td>
<td>$\rho(l_{13}, l_{22}+l_{24})$</td>
</tr>
<tr>
<td>$D_{23}$</td>
<td>Right corner of right eye to right mouth corner</td>
<td>$\rho(l_{10}, l_{24})$</td>
</tr>
<tr>
<td>$D_{24}$</td>
<td>Upper side of left eye to right corner of left eyebrow</td>
<td>$\rho(l_{14}+l_{15}, l_{17})$</td>
</tr>
<tr>
<td>$D_{25}$</td>
<td>Upper side of right eye to left corner of right eyebrow</td>
<td>$\rho(l_{14}+l_{15}, l_{14})$</td>
</tr>
<tr>
<td>$D_{26}$</td>
<td>Upper side of left eye to left corner of right eyebrow</td>
<td>$\rho(l_{14}+l_{15}, l_{14})$</td>
</tr>
</tbody>
</table>

![Fig. 2. Subset of landmarks extracted by the Asthana et al. algorithm [2] used in the proposed algorithm.](image)
TABLE III. EXTRACTED DYNAMIC FEATURE STATISTICS. \( \eta \) DENOTES THE NUMBER OF FRAMES, \( D \) DENOTES THE RESPECTIVE DYNAMIC FEATURE, \( V(t) = \frac{dO}{dt} \) DENOTES THE SPEED, \( A = \frac{d^2O}{dt^2} \) DENOTES THE ACCELERATION, \( \omega \) DENOTES THE FRAME RATE OF THE VIDEO SEQUENCE. THE SUPERSCRIPT + DENOTES THE ONSET, SUPERSCRIPT a DENOTES THE APEX, SUPERSCRIPT − DENOTES THE OFFSET.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>( \frac{\mu(D^+)}{\eta(D^+)} )</td>
</tr>
<tr>
<td>Duration Ratio</td>
<td>( \frac{\mu(D^+)}{\eta(D^+)} )</td>
</tr>
<tr>
<td>Maximal Amplitude</td>
<td>( \max(D) )</td>
</tr>
<tr>
<td>STD of Amplitude</td>
<td>( \text{std}(D) )</td>
</tr>
<tr>
<td>Mean Amplitude</td>
<td>( \text{mean}(D^+) )</td>
</tr>
<tr>
<td>Total Amplitude</td>
<td>( \sum(D^+) )</td>
</tr>
<tr>
<td>Net Amplitude</td>
<td>( \sum(D^+) - \sum(</td>
</tr>
<tr>
<td>Amplitude Ratio</td>
<td>( \frac{\sum(D^+)}{\sum(D^+)+\sum(D^-)} )</td>
</tr>
<tr>
<td>Maximal Speed</td>
<td>( \max(V^+) )</td>
</tr>
<tr>
<td>Mean Speed</td>
<td>( \text{mean}(V^+) )</td>
</tr>
<tr>
<td>Maximum Acceleration</td>
<td>( \max(A^+) )</td>
</tr>
<tr>
<td>Mean Acceleration</td>
<td>( \text{mean}(A^+) )</td>
</tr>
<tr>
<td>Net Ampl., Duration Ratio</td>
<td>( \frac{\sum(D^+) - \sum(</td>
</tr>
</tbody>
</table>

The most relevant information:

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

with \( I \) being the mutual information function, \( c \) the target class, \( F \) the feature set, and \( S_{m-1} \) set of \( m-1 \) features. The mutual information \( I \) of a feature \( f_j \) and the target class \( c \) is computed based on the related probability density functions \( p(f_j), p(c) \) and \( p(f_j, c) \) as follows

\[
I(f_j; c) = \int \int p(f_j, c) \log \frac{p(f_j, c)}{p(f_j)p(c)} df_j dc.
\]

D. Classification

A pattern classifier, trained on labeled data, is used to classify the feature vector into one of two classes: male or female.

We utilized linear Support Vector Machines (SVM) [15], AdaBoost [6] and Bagged Trees [10] in this work. For SVM the Gaussian RBF kernel is used. The optimum values for C and the kernel parameter \( \gamma \) are obtained by a grid-search of the parameter space based on the training set.

E. Extracted Appearance Features

OpenBR [49] is a publicly available open source software for biometric recognition and evaluation. We utilize the gender estimation algorithm, based on the work of Klare et al. [48]. Specifically, a face image is represented by extracting histograms of local binary pattern (LBP) and scale-invariant feature transform (SIFT) features computed on a dense grid of patches. Subsequently, the histograms from each patch are projected onto a subspace generated using Principal Component Analysis (PCA) in order to obtain a feature vector. Support Vector Machine (SVM) is used for the final gender estimation. The OpenBR gender classification algorithm has been validated on a FERET\textsuperscript{5} subset, attaining accuracies of 96.91% and 82.98% for male and female classification, respectively and an overall true classification rate of 90.57% [17], outperforming other algorithms (Neural Network, Support Vector Machine, etc.) on the same dataset [54].

how-old.net is a website (http://how-old.net/) launched by Microsoft for online age and gender recognition. Images can be uploaded and as an output age and gender labels are provided. The underlying algorithm and training dataset are not publicly disclosed.

Commercial Off-the-Shelf (COTS) is a commercial face detection and recognition software, which includes a gender classification routine. The underlying algorithm and the training dataset that were used are not publicly disclosed. The system does not provide a mechanism to re-train the algorithm based on an external dataset; instead it is a black box that outputs a label (i.e., male or female) along with a confidence value.

Since the video-sequences of the UvA-NEMO dataset start with the neutral expression of the portrayed subject, the first frame is utilized to extract appearance features.

F. Fusion of Dynamic and Appearance Features

We concatenate score-levels obtained from the appearance based-algorithms with features obtained from the feature selection step of the dynamics-framework. We utilize PCA to reduce the dimension and obtain a fused feature vector.

\textsuperscript{5}http://www.nist.gov/itl/iad/ig/colorferet.cfm
III. UvA-NEMO SMILE-DATASET

The UvA-NEMO Smile Dataset, introduced by Dibeklioğlu et al. [28], consists of multiple video sequences of 400 subjects (185 females, 215 male). The age of the subjects ranges from 8 to 76 years, see Fig. 4 for the age-distribution. For the most of the subjects there are two videos per subject displaying: (a) spontaneous smile and (b) posed smile. To elicit spontaneous smiles, each subject was displayed a short funny video segment. Each video starts and ends with neutral or a near-neutral expression of the subject (see Fig. 5). The pose of the subjects is frontal and the illumination condition is reasonably constant across subjects. The resolution of the videos is 1920 x 1080 pixels at a framerate of 50 frames per second. This dataset has been used for the analysis of smiles for different ages [28] and for smile-based age analysis [27].

We note that the ethnicity of subjects in the UvA-NEMO dataset is predominantly Caucasian, hence the current study does not reflect on covariates such as ethnicity, as well as social and cultural background.

---

6http://www.uva-nemo.org

IV. RESULTS

In order to evaluate the performance of the proposed gender estimation algorithm, we employ a 15-fold cross-validation scheme. Here, the UvA-NEMO dataset is divided into 15 folds with approximately 24 subjects in each fold. 14 folds are used for training the dynamic gender-estimation algorithm, and the remaining fold is used for testing it. This is repeated 15 times and reported results are the average thereof. Note that the subjects in the training set are not present in the test set.
A. Dynamics versus Appearance

Table IV firstly depicts the discriminative power of the two complementary characteristics individually for spontaneous smiles. As mentioned above, we report age-based gender recognition accuracy. Since training is required for the dynamics-based gender estimation (and hence larger amount of subjects per group), we merge age-groups to two main groups: < 20 years and > 19 years and provide the associated results in Table IV. We observe that the appearance-based gender algorithms perform significantly better for the age category > 19 years and rather poorly in the age category < 20 years. This is due to age-unbalanced training sets or merely due to poor feature performance for toddlers and adolescents, due to low sexual dimorphism. The related confusion matrices for the age category > 19 years are shown in Table V.

Dynamics-based gender estimation: Interestingly, dynamic features (True Gender Classification Rate $TGCR = 59.44\%$) outperform two of the three appearance-based features ($TGCR_{OpenBR} = 52.45\%$ and $TGCR_{how-old.net} = 51.05\%$) in the first age-category. While, appearance-based features are more reliable for the age category > 19 years with $TGCR_{OpenBR} = 78.04\%$, $TGCR_{how-old.net} = 93.46\%$, $TGCR_{COTS} = 92.52\%$; dynamics-based features obtain a noticeable accuracy of 67.81%. The latter suggests that facial smile-dynamics carry substantial cues related to gender of the subject. The confusion matrix is rather balanced in the dynamics-based gender estimation (Table V (d)).

We note that fusion of appearance and smile-dynamics-based gender estimation either increases the performance of appearance-based algorithms (e.g., for OpenBR in both age classes, for how-old.net in the younger age-class and for COTS in the older age-class) or does not impact it negatively. Related confusion matrices are shown in Table V.

In our related work [9], we have presented a holistic approach for smile-based gender estimation, that extracts spatiotemporal features based on dense trajectories, represented by a set of descriptors encoded by Fisher Vectors. The associated true gender classification rates account for 86.3% for adolescents, and 91.01% for adults.

B. Spontaneous versus posed smile

We also provide results on the posed-smile subset of the UvA-NEMO dataset presented in Table VI. Interestingly, the associated dynamics-based gender-estimation accuracy resembles strongly the spontaneous-smile-case. The difference in performance origins in the slightly larger posed-smile subset-size, that contributes to larger training-sets in the case of dynamics-based gender classification, as well as in the fusion of appearance and dynamics-based features. Nevertheless, the results suggest that dynamics of posed smiles carry significant cues on gender, similarly to spontaneous smiles. The related confusion matrices are shown in Table VII.

This result is in agreement with psychological findings, that show that females are more accurate expressers of emotion, when posing deliberately and when observed unobtrusively [11], hinting that posing a smile carries gender-specific cues.

C. Gender divergence in spontaneous and posed smiles

We seek to answer the question, whether males or females pose smiles more genuinely and whether there is a significant divergence. Towards this, we combine features in all possible sets and compute Euclidean distances between sets in the spontaneous and the associated sets in the posed-smile-case. Fig. 6 illustrates the related results for the most diverging case between males and females. Females have slightly lower distances, suggesting that females pose smiles more realistically; however, the disparity is not significant. This tendency conforms with previous psychological findings [11].

![Fig. 6. Distributions of Euclidean distances between posed and spontaneous feature vectors for male and female subjects in the UvA-NEMO dataset.](image)

D. Discriminative Features

We here analyze the individual discriminability of the selected dynamic-features for the 27 distances. Towards this, we estimate gender based on each feature individually. Hence, we train and test an SVM-classifier with each feature individually for the two age groups, < 20 and > 19 years. We report for each age group the most discriminative features respectively (see Table VIII and Table IX). The most striking outcome is that the majority of discriminative features are in the mouth region. It is also interesting to note that while for the younger group D10 (Center of mouth to right mouth corner) and D7 (Center of mouth to left side of upper lip) and the onset-phase are predominant, for the older group D3 (Length of mouth) and mainly the offset-phase is more profound. This hints that
TABLE IV. SPONTANEOUS SMILE. TRUE GENDER CLASSIFICATION RATES. AGE GIVEN IN YEARS.

<table>
<thead>
<tr>
<th>Age</th>
<th>&lt; 10</th>
<th>10 – 19</th>
<th>20 – 29</th>
<th>30 – 39</th>
<th>40 – 49</th>
<th>&gt; 49</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj. amount</td>
<td>48</td>
<td>95</td>
<td>60</td>
<td>49</td>
<td>72</td>
<td>33</td>
</tr>
<tr>
<td>OpenBR</td>
<td>58.33%</td>
<td>50.53%</td>
<td>81.67%</td>
<td>75.51%</td>
<td>75%</td>
<td>81.82%</td>
</tr>
<tr>
<td>how-old.net</td>
<td>39.58%</td>
<td>56.84%</td>
<td>95%</td>
<td>87.76%</td>
<td>98.61%</td>
<td>87.88%</td>
</tr>
<tr>
<td>COTS</td>
<td>77.68%</td>
<td>76.84%</td>
<td>93.33%</td>
<td>89.8%</td>
<td>94.44%</td>
<td>90.91%</td>
</tr>
</tbody>
</table>

Merger Age-Groups

<table>
<thead>
<tr>
<th>Subj. amount</th>
<th>&lt; 20</th>
<th>&gt; 19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamics (PCA, SVM)</td>
<td>59.44%</td>
<td>67.81%</td>
</tr>
<tr>
<td>OpenBR</td>
<td>52.45%</td>
<td>78.04%</td>
</tr>
<tr>
<td>how-old.net</td>
<td>51.05%</td>
<td>93.46%</td>
</tr>
<tr>
<td>how-old.net + Dynamics (Tree)</td>
<td>60.8%</td>
<td>93.46%</td>
</tr>
<tr>
<td>COTS</td>
<td>76.92%</td>
<td>92.52%</td>
</tr>
<tr>
<td>COTS + Dynamics (Tree)</td>
<td>76.92%</td>
<td>93%</td>
</tr>
</tbody>
</table>

TABLE V. SPONTANEOUS SMILE IN AGE CATEGORY > 19: CONFUSION MATRIX FOR MALES AND FEMALES FOR (A) APPEARANCE FEATURES #1 (OpenBR) (denoted as App. 1), (B) APPEARANCE FEATURES #2 (how-old.net) (denoted as App. 2), (C) APPEARANCE FEATURES #3 (COTS) (denoted as App. 3), (D) DYNAMIC FEATURES (denoted as Dyn.), (E) DYNAMIC AND APPEARANCE FEATURES #1 (denoted as Dyn. + App. 1), (F) DYNAMIC AND APPEARANCE FEATURES #2 (denoted as Dyn. + App. 2), (G) DYNAMIC AND APPEARANCE FEATURES #3 (denoted as Dyn. + App. 3).

<table>
<thead>
<tr>
<th>(a) App. 1</th>
<th>(b) App. 2</th>
<th>(c) App. 3</th>
<th>(d) Dynamics</th>
<th>(e) Dyn. + App. 1</th>
<th>(f) Dyn. + App. 2</th>
<th>(g) Dyn. + App. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Male</td>
<td>61.8%</td>
<td>38.2%</td>
<td>94.5%</td>
<td>5.5%</td>
<td>99.1%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Female</td>
<td>4.8%</td>
<td>95.2%</td>
<td>7.7%</td>
<td>92.3%</td>
<td>14.4%</td>
<td>85.6%</td>
</tr>
</tbody>
</table>

TABLE VI. SPONTANEOUS SMILE IN AGE CATEGORY > 19: CONFUSION MATRIX FOR MALES AND FEMALES FOR (A) APPEARANCE FEATURES #1 (OpenBR) (denoted as App. 1), (B) APPEARANCE FEATURES #2 (how-old.net) (denoted as App. 2), (C) APPEARANCE FEATURES #3 (COTS) (denoted as App. 3), (D) DYNAMIC FEATURES (denoted as Dyn.), (E) DYNAMIC AND APPEARANCE FEATURES #1 (denoted as Dyn. + App. 1), (F) DYNAMIC AND APPEARANCE FEATURES #2 (denoted as Dyn. + App. 2), (G) DYNAMIC AND APPEARANCE FEATURES #3 (denoted as Dyn. + App. 3).

<table>
<thead>
<tr>
<th>(a) App. 1</th>
<th>(b) App. 2</th>
<th>(c) App. 3</th>
<th>(d) Dynamics</th>
<th>(e) Dyn. + App. 1</th>
<th>(f) Dyn. + App. 2</th>
<th>(g) Dyn. + App. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Male</td>
<td>62.34%</td>
<td>37.66%</td>
<td>94.81%</td>
<td>5.19%</td>
<td>98.7%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Female</td>
<td>1.52%</td>
<td>98.48%</td>
<td>7.58%</td>
<td>92.42%</td>
<td>15.15%</td>
<td>84.85%</td>
</tr>
</tbody>
</table>

... sexual dimorphism can be gleaned from the asymmetrical-onset in adolescents. On a related note, a recent psychological study [13] has found that expressions shown on the left hemi-face (LHF) were rated as more intense, and furthermore that spontaneous expressions start earlier in the LHF. Hence expressions in both hemi-faces are not fully redundant.

Description of most discriminative features In adolescents, females tended to show longer Duration Ratio – Offset and longer Duration – Onset on the right side of the mouth and higher Amplitude Ratio – Onset on the left side of the mouth, than males. In adults, females tended to show higher Mean Amplitude – Apex of mouth opening, higher Maximum Amplitude on the right side of the mouth, as well as faster Mean Speed – Offset on the left side of the mouth, than males. Figure 7 illustrates the boxplots for the five most discriminative features in the age category > 19 years for spontaneous smile.

We here note, that the selected features for the proposed algorithm in previous sections do not correspond to the presented features in this section, since a mutual information function prunes out correlated features in the selection process, which we do not consider here.

V. Conclusions

In this work we introduced smile-based dynamic facial feature extraction for gender estimation. The proposed dynamics-based gender estimation algorithm predominantly improves the performance of three state-of-the-art appearance-based gender estimation algorithms. We observe that dynamics can outperform appearance-based features for subjects younger than 20 years old; while facial appearance features are more discriminative for older subjects. We show that appearance and dynamics-based features are complementary and the combination thereof beneficial. Our results further suggest that gender...
is mainly exhibited in dynamics in the mouth-region among the studied facial dynamic-features. Finally, we analyzed the gender-dimorphism of both, spontaneous and posed smiles and observe that both carry substantial cues for gender.

### APPENDIX A

**ACKNOWLEDGMENT**

This work was carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement nr. 246016.

**REFERENCES**


Table VIII. **Most Discriminate Dynamic Features for Age < 20.** TGCR...True Gender Classification Rate.

<table>
<thead>
<tr>
<th>Distance</th>
<th>Feature</th>
<th>TGCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>D10</td>
<td>Duration Ratio Offset</td>
<td>65.73%</td>
</tr>
<tr>
<td>D7</td>
<td>Amplitude Ratio Onset</td>
<td>63.64%</td>
</tr>
<tr>
<td>D10</td>
<td>Duration Onset</td>
<td>62.94%</td>
</tr>
<tr>
<td>D10</td>
<td>Total Amplitude Onset</td>
<td>62.24%</td>
</tr>
<tr>
<td>D9</td>
<td>Maximum Amplitude</td>
<td>62.24%</td>
</tr>
<tr>
<td>D10</td>
<td>Mean Amplitude Offset</td>
<td>62.24%</td>
</tr>
<tr>
<td>D7</td>
<td>Amplitude Ratio Offset</td>
<td>62.24%</td>
</tr>
<tr>
<td>D10</td>
<td>Maximum Amplitude</td>
<td>61.54%</td>
</tr>
</tbody>
</table>

Table IX. **Most Discriminate Dynamic Features for Age > 19.** TGCR...True Gender Classification Rate.

<table>
<thead>
<tr>
<th>Distance</th>
<th>Feature</th>
<th>TGCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>D11</td>
<td>Mean Amplitude Apex</td>
<td>62.15%</td>
</tr>
<tr>
<td>D9</td>
<td>Maximum Amplitude</td>
<td>61.68%</td>
</tr>
<tr>
<td>D9</td>
<td>Mean Speed Offset</td>
<td>61.54%</td>
</tr>
<tr>
<td>D5</td>
<td>Mean Acceleration Offset</td>
<td>60.28%</td>
</tr>
<tr>
<td>D5</td>
<td>Amplitude Ratio Offset</td>
<td>60.28%</td>
</tr>
<tr>
<td>D5</td>
<td>Duration Offset</td>
<td>60.28%</td>
</tr>
<tr>
<td>D5</td>
<td>Maximum Acceleration Offset</td>
<td>60.14%</td>
</tr>
<tr>
<td>D5</td>
<td>Duration Apex</td>
<td>59.81%</td>
</tr>
</tbody>
</table>


[42] U. Hess and P. Thibault. Why the same expression may not mean the same when shown on different faces or seen by different people. Affective information processing, pages 145–158, 2009.


