Attended End-to-End Architecture for Age Estimation From Facial Expression Videos

Wenjie Pei, Hamdi Dibeklioğlu, *Member, IEEE*, Tadas Baltrušaitis, and David M. J. Tax

***Abstract*—The main challenges of age estimation from facial expression videos lie not only in the modeling of the static facial appearance, but also in the capturing of the temporal facial dynamics. Traditional techniques to this problem focus on constructing handcrafted features to explore the discriminative information contained in facial appearance and dynamics separately. This relies on sophisticated feature-refinement and framework-design. In this paper, we present an end-to-end architecture for age estimation, called Spatially-Indexed Attention Model (SIAM), which is able to simultaneously learn both the appearance and dynamics of age from raw videos of facial expressions. Specifically, we employ convolutional neural networks to extract effective latent appearance representations and feed them into recurrent networks to model the temporal dynamics. More importantly, we propose to leverage attention models for salience detection in both the spatial domain for each single image and the temporal domain for the whole video as well. We design a specific spatially-indexed attention mechanism among the convolutional layers to extract the salient facial regions in each individual image, and a temporal attention layer to assign attention weights to each frame. This two-pronged approach not only improves the performance by allowing the model to focus on informative frames and facial areas, but it also offers an interpretable correspondence between the spatial facial regions as well as temporal frames, and the task of age estimation. We demonstrate the strong performance of our model in experiments on a large, gender-balanced database with 400 subjects with ages spanning from 8 to 76 years. Experiments reveal that our model exhibits significant superiority over the state-of-the-art methods given sufficient training data.**

***Index Terms*—Age estimation, end-to-end, attention, facial dynamics.**

I. Introduction

HUMAN age estimation from faces is an important research topic due to its extensive applications ranging from surveillance monitoring [1], [2] to forensic art [3], [4] and social networks [5], [6]. The widely-used discriminative features for age estimation are appearance-related, such as wrinkles in the face, skin texture and luster, hence plenty of prevalent methods focus on modeling the appearance information from the static face [7], [8]. Recent studies [9], [10] also indicate that the dynamic information in facial expressions like temporal properties of a smile can be leveraged to significantly improve the performance of age estimation. It is reasonable since there are intuitive temporal dynamics involved in facial expressions which are relevant to the age. For instance, the exhibited facial movement like wrinkles in the smiling process is different for people of different ages.

The traditional approaches to age estimation from facial expression videos focus on constructing handcrafted features to explore the discriminative information contained in static appearance and temporal dynamics separately, and then combining them into an integrated system [9], [10]. However, these kinds of methods rely on sophisticated feature design. In this work, we propose a novel end-to-end architecture for age estimation from facial expression videos, which is able to automatically learn the static appearance in each single image and the temporal dynamics contained in the facial expression simultaneously. In particular, we employ convolutional neural networks (CNNs) to model the static appearance due to its successful representation learning in the image domain. The learned latent appearance features for each image are subsequently fed into recurrent networks to model the temporal dynamics. In this way, both static appearance and temporal dynamics can be integrated seamlessly in an end-to-end manner. A key benefit of this design is that the learned static appearance features and temporal dynamic features are optimized jointly, which can lead to better performance for the final task of age estimation. Additionally, the end-to-end manner of our method avoids separating feature design from the age regression as a two-stage procedure, which is the typical way of the methods using handcrafted features.

Attention models have been proposed to let models learn by themselves to pay attention to specific regions in an image [11] or different segments in a sequence [12], [13] according to the relevance to the aimed task. Likewise, different facial parts (in each single image) and different phases of the expression (in the inspected video) may exhibit varying degrees of salience to age estimation. To incorporate attention, we customize a specific attention module for spatial facial salience detection. To detect temporal salience in the sequential expression, we mount a temporal attention module upon the recurrent networks. It serves as a filtering layer to determine the amount of information of each frame to be incorporated into final age regression task. All functional modules of our proposed Spatially-Indexed Attention

Manuscript received November 22, 2017; revised October 30, 2018 and August 20, 2019; accepted October 8, 2019. Date of publication October 24, 2019; date of current version November 27, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Baoxin Li. *(Corresponding author: Wenjie Pei.)*

W. Pei is with the Harbin Institute of Technology, Shenzhen 518055, China (e-mail: wenjiecoder@gmail.com).

H. Dibeklioğlu is with the Department of Computer Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: dibeklioglu@cs.bilkent.edu.tr).

T. Baltrušaitis is with Microsoft Corporation, Microsoft Research Lab Cambridge, Cambridge CB1 2FD, U.K. (e-mail: tabaltru@microsoft.com).

D. M. J. Tax is with the Pattern Recognition Laboratory, Delft University of Technology, 2628 CD, Delft, The Netherlands (e-mail: d.m.j.tax@tudelft.nl).

Digital Object Identifier 10.1109/TIP.2019.2948288

Model (SIAM) including convolutional networks for learning appearance, recurrent networks for learning dynamics, two attention modules as well as the final age regression module can be trained jointly, without any manual intervention.

Extensive experiments on a real-world database demonstrate the substantial superiority of our model over the state-of-the-art methods. Our model has the capacity to learn and it could do even better on more data while other models potentially saturate and do not get better no matter how much data you give them. Notably, larger training data tends to explore more potential of our model and expand its advantages compared to other methods.

II. Related Work

In this study, we propose to effectively learn spatial and temporal patterns of aging in an attended end-to-end manner for a more reliable age estimation. To comprehend the related concepts, in this section, the literature on automatic age estimation will be summarized, and an overview of neural attention models will be given.

A. Automatic Age Estimation

Based on the fact that age information is crucial to understand requirements or preferences of people, automatic estimation of age from face images has quite a few real-life applications, and thus, it has been a very active area of research over the last two decades. Biometric search/authentication in facial image databases, denying purchase of unhealthful products (e.g. alcohol and tobacco) by underage customers, and personalization/adaptation of interactive systems to the users (displaying personalized advertisements) can be counted as some examples of the use of age estimation.

One of the major requirements in facial age estimation is the capturing/modeling of facial properties that change by age. These patterns are related to craniofacial development [14] and alteration in facial skin (e.g. wrinkles) [15]. While the majority of earlier studies in the area focus on describing such patterns using engineered (handcrafted) representations such as local binary patterns (LBP) [16] and its variations [17], Gabor filter based biologically-inspired aging features (BIF) [18], and shape-based features [19], some studies propose to capture aging patterns through learning-based approaches such as subspace learning [1], [20]–[23], PCA-tree-based encoding [24], and metric learning [25], [26].

The increase in the number and size of facial age databases, and the recent dramatic improvements in the field of deep learning have shifted the focus towards deep architectures to learn complex (nonlinear) patterns of facial aging from a large collection of face images. For instance, [27] presents the first exploration of employing CNNs for age estimation, where representations obtained from different layers of CNN are used. In [28], a multi-task CNN architecture is proposed to optimize the facial representation jointly for age and gender estimation. Reference [29] models the features extracted from a pre-trained CNN (VGG-Face [30]) using the kernel extreme learning machines. Reference [31] uses VGG-16 CNN architecture [32] (pre-trained on ImageNet images) and fine-tunes the model using a multi-class classifier for age estimation. Then, softmax-normalized output probabilities are used for the final prediction. Differently from conventional methods, [31] solely employs face detection for face alignment rather than using facial landmark detection, leading to a more accurate age estimation. In [33], Agustsson *et al.* complement [31] with their proposed Anchored Regression Network (rather than employing softmax classifier on top the VGG-16 CNN architecture), enhancing the reliability. In a recent study [34], Xing *et al.* have analyzed different loss functions and CNN architectures for age estimation as well as employing a joint optimization together with race and gender classification tasks. Unlike previous methods that use deep neural networks in a supervised manner, [35] investigates an unsupervised learning framework for CNN features by applying k-means to learn convolutional filters of a single-layer CNN. Obtained CNN features are refined by subsequent unsupervised recurrent neural networks (with randomly initialized parameters) and projected into a discriminative subspace for age estimation by training an SVM or SVR.

In contrast to regression and multi-class classification, some studies approach age estimation as an ordinal ranking problem. For instance, [36] presents a deep (category-based) ranking model that combines deep scattering transform and ordinal ranking. Reference [37] formulates the problem as ordinal regression using a series of binary classification tasks which are jointly optimized by a multiple output CNN architecture. In [38], instead of a multiple output model, a series of basic CNNs are employed (a separate CNN for each ordinal age group), and their binary outputs are aggregated. Such an approach allows capturing different patterns for different age groups.

The use of deep architectures has significantly improved the reliability of automatic age estimation, especially under pose and illumination variations. Facial expression variation, however, is still a challenge since expressions form deformations on the facial surface that can be confused with aging-related wrinkles. Yet, only a few recent works in the literature explore solutions for this issue [39]–[42]. Guo and Wang [39] model correlation between aging features of the neutral face and a specific expression (i.e. smile) of individuals. Learned correlation is used to map the features of expressive faces to those of neutral ones. In this way, the confusing influence of expressions in aging patterns are removed. However, this method requires an accurate estimation of facial expressions (before the age estimation), and a separate training for each expression of interest using neutral and expressive face images of each subject in the training dataset. Reference [40] learns a common subspace for a set of facial expressions that reduce the influence of expressions while preserving the aging patterns. Reference [41] defines four age groups, and models each facial expression in each age group as a different class for cross-expression age estimation. In a similar manner, [42] proposes a multi-task framework that jointly learns the age and expression using the latent structured support vector machines.

Interestingly, until a recent work of Hadid [43], none of the methods in the literature have used facial videos for age estimation. In [43], the volume local binary patterns (VLBP)

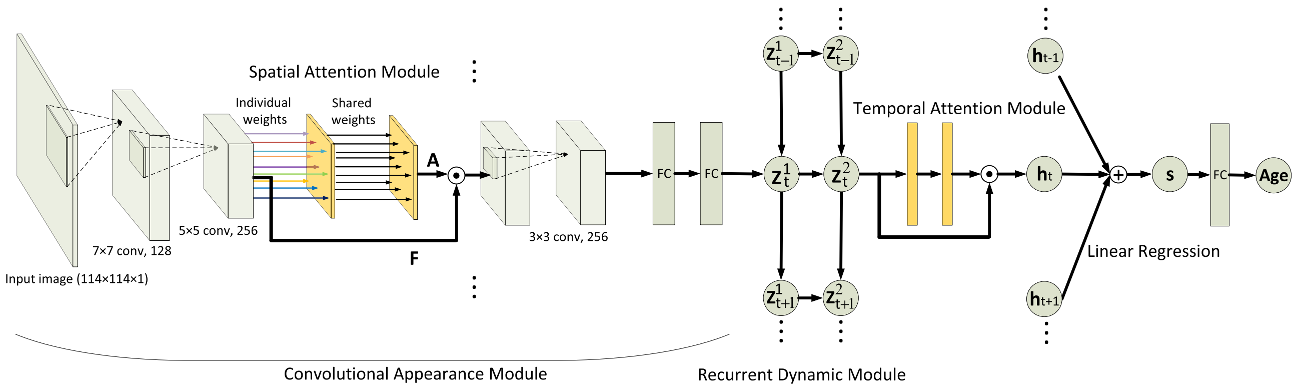


Fig. 1. The graphical representation of our Spatially-Indexed Attention Model (SIAM) for age estimation. Note that we only give one instance of the convolutional appearance module, the spatial attention module and temporal attention module for one frame at time step t. Other frames share exactly the same structure for all three modules.

features are employed to describe spatio-temporal information in videos of talking faces for age grouping (child, youth, adult, middle-age, and elderly). Yet, this video-based method ([43]) could not perform as accurate as the image-based methods. On the other method, more recently, Dibeklioğlu *et al.* have presented the first successful example of video-based age estimation [9], where displacement characteristics of facial landmarks are represented by a set of handcrafted descriptors extracted from smile videos, and combined with spatio-temporal appearance features. In a follow-up study, Dibeklioğlu *et al.* [10] have enhanced their engineered features so as to capture temporal changes in 3D surface deformations, leading to a more reliable age estimation. It is important to note that these two studies exploit the aging characteristics hidden in temporal dynamics of facial expressions, rather than reducing the influence of expressions in aging features. Thus, following the informativeness of temporal dynamics of facial expressions and based on the success of deep models, the current study proposes a deep temporal architecture for automatic age estimation.

Because the literature on automatic age estimation is extensive, for further information, we refer the reader to [4], [44], and to the more recent [45].

B. Attention Models

Much of progress in neural networks was enabled by so called neural attention, which allows the network to focus on certain elements of a sequence [12], [13], [46] or certain regions of an image [11] when performing a prediction. The appeal of such models comes from their end-to-end nature, allowing the network to learn how to attend to or align data before making a prediction.

It has been particularly popular in encoder-decoder frameworks, where it was first introduced to better translate between languages [13]. The attention network learned to focus on particular words or phrases when translating sentences, showing large performance gains on especially long sequences. It has also been extensively used for visual image and video captioning, allowing the decoder module to focus on parts of the image it was describing [11]. Similarly, the neural attention models have been used in visual question answering tasks, helping the alignment between words in the question and regions in the image [47]. Spatial Transformer Networks which focus on a particular area of image can also be seen as a special case of attention [48]. Somewhat relatedly, work in facial expression analysis has explored using particular regions for facial action unit detection [49], [50], however, they did not explore dynamically attending to regions depending on the current facial appearance. Our work is inspired by these attention models, but explores different ways of constructing neural attention and applying it to age estimation.

III. Method

Given a video displaying the facial expression of a subject, the aim is to estimate the age of that person. Next to that, the model is expected to capture the salient facial regions in the spatial domain and the salient phase during facial expression in the temporal domain. Our proposed Spatially-Indexed Attention Model (SIAM) is composed of four functional modules: 1) a convolutional appearance module for appearance modeling, 2) a spatial attention module for spatial (facial) salience detection, 3) a recurrent dynamic module for facial dynamics, and 4) a temporal attention module for discriminating temporal salient frames. The proposed model is illustrated in Figure 1. We will elaborate on the four modules in a bottom-up fashion and explain step by step how they are integrated into an end-to-end trainable system.

A. Convolutional Appearance Module

Convolutional neural networks (CNNs) have achieved great success for automatic latent representation learning in the image domain. We propose to employ CNNs to model the static appearance for each image of the given video. Compared to conventional handcrafted image features which are generally designed independently of the aimed task, the features learned automatically by CNNs tend to describe the aimed task more accurately due to the parameters learning by back-propagation from the loss function of the aimed task.

Our model contains three convolutional layers with coarse-to-fine filters and subsequent two fully-connected layers.

The output of the last fully-connected layer is fed as input to recurrent modules at the corresponding frame. Response-normalization layers [51] and Max-pooling layers follow the first and second convolutional layers. The ReLU [52] function is used as the nonlinear activation function for each convolutional layer as well as the fully-connected layer. We found that creating deeper networks did not lead to a performance improvement, possibly due to comparatively small training dataset used. It should be noted that the spatial attention module presented subsequently in Section III-B is embedded among convolutional layers to perform facial salience detection. The details are depicted in Figure 1.

For each frame, the input image with size $114 \times 114 \times 1$ is filtered by the first convolutional layer with 128 kernels with size 7×7 and stride size 2. The second convolutional layer filters the output of the first convolutional layer with 256 kernels of size 5×5. The spatial attention module (Section III-B) then takes as input the output of the second convolutional layer and extracts the salient regions in the face. The generated attended feature map containing salience is filtered by the third convolutional layer that has 256 kernels of size 3×3. The first fully-connected layer has 4096 neurons while the second fully-connected layer has the same number of neurons as the subsequent recurrent network (as will be explained in Section III-C).

It should be noted that the same convolutional module is shared across all frames in the time domain. Thus, the forward pass of the convolutional module can be computed in parallel for all frames. In the backward pass, the parameters in the convolutional module are optimized by back-propagating output gradients of the upper recurrent module through all frames.

B. Spatial Attention Module

The goal of the spatial attention module is to dynamically estimate the salience and relevance of different image portions for the downstream task (in our case age estimation). It is implemented as a feature map filter embedded after one of the convolutional layers in the convolutional appearance modules to preserve the information based on the calculated saliency score.

Formally, suppose the output volume $\mathbf{F}$ of a convolutional layer L has dimensions $M \times N \times C$, with C feature maps of size $M \times N$. The spatial attention module embedded after the convolutional layer L is denoted by a matrix $\mathbf{A}$ with the same size $M \times N$ as the feature map. The element $\mathbf{A}_{ij}$ of $\mathbf{A}$ indicates the attention weight (interpreted as saliency score) for the feature vector $\mathbf{F}_{ij}$ composed of C channels located at (i, j) (i.e., $|\mathbf{F}_{ij}| \equiv C$) in the feature map. Each feature vector corresponds to a certain part of the input image (i.e., receptive field). Therefore, the receptive field of the attention becomes larger when the attention module is inserted in latter convolutional layers. Section V-A.2 presents an experimental comparison of the different positions of spatial attention module in the convolutional appearance module. In practice, we insert the spatial attention module after the second convolutional layer.

We propose a spatially-indexed attention mechanism to model $\mathbf{A}$. Concretely, the attention weight $\mathbf{A}$ is modeled by two fully-connected layers: the first layer is parameterized by individual weights for each entry of feature map while the second layer shares the transformation weights across the whole map. Thus the attention weight $\mathbf{A}_{ij}$ is modeled as:

$$\mathbf{A}_{ij} = \sigma(\mathbf{u}^{\top} \tanh(\mathbf{W}_{ij}\mathbf{F}_{ij} + \mathbf{b}_{ij}) + c) \tag{1}$$

Herein, $\mathbf{W}_{ij} \in \mathbb{R}^{d \times C}$ is the transformation matrix for the first fully-connected layer and $\mathbf{u}^{\top} \in \mathbb{R}^{d}$ is the weight vector for the second fully-connected layer to fuse the information from different channels and c is a bias term. A sigmoid function σ is employed as the activation function at the top layer of the attention module to constraint the attention weight to lie in the interval $[0, 1]$. The obtained attention matrix $\mathbf{A}$ controls the information flowing into the subsequent layer in the convolutional appearance module by an element-wise multiplication to each channel (feature map) of F:

$$\mathbf{I} = \mathbf{F} \odot \mathbf{A} \tag{2}$$

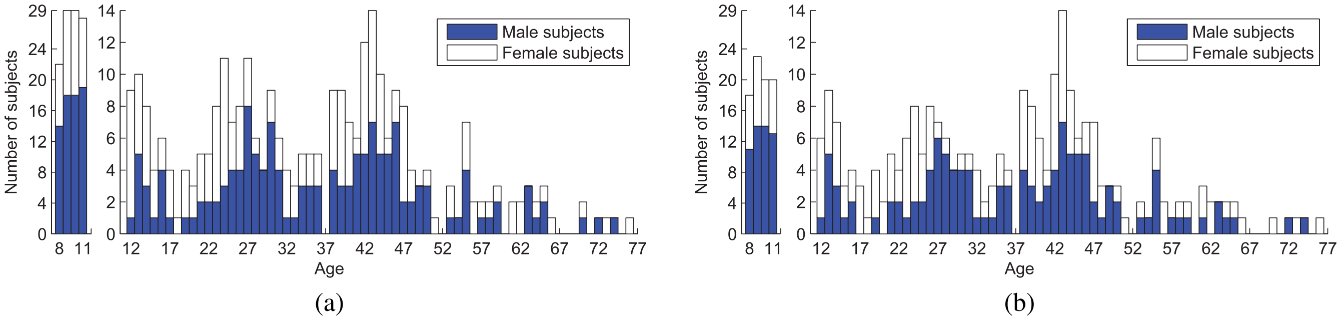
Here $\mathbf{I}$ is the output of the spatial attention module, which is fed into the subsequent layer of the convolutional appearance module.

It is worth mentioning that we use individual weights ($\mathbf{W}_{ij}$) for the first layer and shared weights $\mathbf{u}$ for the second fusion layer in the spatial attention module (this will be called a spatially-indexed mechanism). The first layer is expected to capture the local detailed (fine-grained) variation while the second fusion layer is designed to capture the global variation and smooth the attention distribution. It is different from the design of the soft attention model for image caption generation [11], in which the transformation weights are shared in both two layers of the attention model. In that scenario, the attention model is used to capture the related objects in the input image to each word of generated caption and the objects are always easily separable from the background scene in the image. By contrast, we aim to capture the salient parts in a face image, which requires to model more detailed variation. Employing shared weights in both layers tends to blur the spatial variation. Besides, the typical attention model is translation-invariant. Namely, if the picture is rearranged, the attention would be very similar, whereas our attention is spatially-indexed. Section V-A.1 provides an comparison between different attention mechanisms by visualizing the learned attention weight distribution.

C. Recurrent Dynamic Module

The temporal facial dynamics are expected to contribute to age estimation, which has been demonstrated by Dibeklioğlu *et al.* [10]. In contrast to the handcrafted dynamics features they use [10], we propose to employ recurrent networks to capture the underlying temporal information automatically. The potential advantages of using recurrent networks are that (1) they learn relevant dynamics feature to the aimed task (age estimation) smoothly and progressively over time; (2) all modules in our model can be trained jointly in an end-to-end manner to be compatible with each other.

Suppose the output appearance feature of last fully-connected layer of convolutional appearance module is



(a)

(b)

Fig. 2. Age and gender distributions of the subjects in (a) the UvA-NEMO Smile and (b) the UvA-NEMO Disgust databases.

$\mathbf{p}_t$ at frame t, then the hidden representation $\mathbf{z}_t$ is calculated by:

$$\mathbf{z}_t = g(\mathbf{W}\mathbf{p}_t + \mathbf{V}\mathbf{z}_{t-1} + b) \tag{3}$$

Herein $\mathbf{W}$ and $\mathbf{V}$ are the transformation matrices for appearance feature in current frame and the hidden representation in the previous frame. We use a ReLU function as the activation function g since it eliminates potential vanishing-gradient problems. In practice, we employ two-layer recurrent networks in our recurrent dynamic module, which is expected to potentially learn more latent temporal dynamics than single-layer recurrent networks.

D. Temporal Attention Module

The attention scheme can be leveraged not only for the selection of the salient facial regions in the spatial domain, but also for the selection of the salient sequential segments (frames) in the temporal domain. Hence we propose to use a temporal attention module on top of the recurrent dynamic module to capture the temporal salience information. The temporal attention module produces an attention weight as the salience score for each frame, thereby filtering the output information flow from the recurrent dynamic module.

Formally, suppose the output hidden-unit representation of the recurrent dynamic module is $\mathbf{z}_t$ at the frame t, then the temporal attention score e_t is modeled by a two-layer perceptron:

$$e_t = \sigma(\mathbf{v}^\top \tanh(\mathbf{M}\mathbf{z}_t + \mathbf{b}) + c) \tag{4}$$

Here $\mathbf{M} \in \mathbb{R}^{n' \times n}$ is the weight matrix and $\mathbf{b}$ is the bias term for the first perceptron layer, $\mathbf{v} \in \mathbb{R}^{n'}$ is the fusion vector of the second layer. Here n' is a hyper-parameter that is the dimension of transformed mid-representation. Again, we use a sigmoid function to constrain the values between 0 and 1. We employ this perceptron to measure the relevance of each frame to the objective task, i.e., age estimation. Next, the attention score is normalized over the whole video to get the final temporal attention weight o_t:

$$o_t = \frac{e_t}{\sum_{t'=1}^{T} e_{t'}} \tag{5}$$

The obtained temporal attention weights are used to control how much information for each frame is taken into account to perform the age estimation. Concretely, we calculate the weighted sum of the hidden-unit representation for all frames of the recurrent dynamic module to be the information summary $\mathbf{s}$ for the whole video:

$$\mathbf{s} = \sum_{t=1}^{T} o_t \mathbf{z}_t \tag{6}$$

Ultimately, the predicted age of the corresponding subject involved in the video is estimated by a linear regressor:

$$\tilde{y} = \mathbf{k} \cdot \mathbf{s} + b \tag{7}$$

where $\mathbf{k}$ contains the regression weights.

E. End-to-End Parameter Learning

Given a training dataset $\mathcal{D} = \{\mathbf{x}^{(n)}_{1,\ldots,T_{(n)}}, y^{(n)}\}_{n=1,\ldots,N}$ containing N pairs of facial videos and their associated subject's age, we learn the involved parameters of all four modules (convolutional appearance module, spatial attention module, recurrent dynamic module, and temporal attention module) and the final linear regressor jointly by minimizing the mean absolute error loss of the training data:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} |\tilde{y}^{(n)} - y^{(n)}| \tag{8}$$

Since all modules and the above loss function are analytically differentiable, our whole model can be readily trained in an end-to-end manner. The loss is back-propagated through four modules successively using back-propagation through time algorithm [53] in the recurrent dynamic module and normal back-propagation way in other parts.

IV. Experimental Setup

A. Datasets

1) UvA-NEMO Smile Database: We evaluate the performance of our proposed age estimation architecture on the UvA-NEMO Smile Database, which was collected to analyze the temporal dynamics of spontaneous/pose smiles for different ages [54]. The database is composed of 1240 smile videos (597 spontaneous and 643 posed) recorded from 400 subjects (185 female and 215 male). The involved subjects span an age interval ranging from 8 to 76 years. Figure 2(a) presents the age and gender distribution.

To collect posed smiles, each subject was asked to pose a smile as realistically as possible. Spontaneous smile was elicited by short and funny video segments. For each subject, approximately five minutes of recordings were made and the genuine smiles were then segmented. A balanced number of spontaneous and posed smiles are selected and annotated by the consensus of two trained annotators for each subject. Each segment of video starts/ends with neutral or near-neutral expressions.

2) UvA-NEMO Disgust Database: To evaluate the proposed approach for other facial expressions, we use the UvA-NEMO Disgust Database [10] that has been recorded during the acquisition of the UvA-NEMO Smile Database using the same recording/illumination setup. The database consists of the (posed) disgust expressions of 324 subjects (152 female, 172 male), where 313 of these subjects are also included in the 400 subjects of the UvA-NEMO Smile Database. For each subject, one or two posed disgust expressions were selected and annotated by seeking consensus of two trained annotators. Each of the segmented disgust expressions starts and ends with neutral or near-neutral expressions. The resulting database has 518 posed disgust videos. Subjects' ages vary from 8 to 76 years as shown in Figure 2(b).

B. Tracking and Alignment of Faces

To normalize face images in terms of rotation and scale, 68 landmarks on facial boundary (17 points), eyes & eyebrows (22 points), nose (9 points), and mouth (20 points) regions are tracked using a state-of-the-art tracker [55]. The tracker employs an extended version of Constrained Local Neural Fields (CLNF) [56], where individual point distribution and patch expert models are learned for eyes, lips and eyebrows. Detected points by individual models are then fit to a joint point distribution model. To handle pose variations, CLNF employs a 3D latent representation of facial landmarks.

The movement of the tracked landmarks is smoothed by the 4253H-twice method [57] to reduce the tracking noise. Then, each face image (in videos) is warped onto a frontal average face shape using a piecewise linear warping. Notice that the landmark points are in the same location for each of the warped/normalized faces. Such a shape normalization is applied to obtain (pixel-to-pixel) comparable face images regardless of expression or identity variations. The obtained images are cropped around the facial boundary and eyebrows, and scaled so as to have a resolution of 114×114 pixels. Images are then converted to gray scale.

C. Settings

Following the experimental setup of Dibeklioğlu et al. [10], we apply a 10-fold cross-validation testing scheme with the same data split to conduct experiments. There is no subject overlap between folds. Each time one fold is used as test data and the other 9 folds are used to train and validate the model. The parameters are optimized independently of test data. For the recurrent dynamic module, the number of hidden units is tuned by selecting the best configuration from the set $\{128, 256, 512\}$ using a validation set. To prevent over-fitting, we adopt Dropout [58] in both the convolutional networks and the recurrent networks and we augment the loss function with L2-regularization terms. Two dropout values, one for the recurrent dynamic module and one for the convolutional appearance module, are validated from the option set $\{0, 0.1, 0.2, 0.4\}$. The L2-regularization parameter λ is validated from the option set $\{0, 1e^{-4}, 3e^{-4}, 5e^{-4}, 1e^{-3}, 3e^{-3}, 5e^{-3}\}$. We perform gradient descent optimization using RMSprop [59]. The gradients are clipped between -5 and 5 [60] to avoid potential gradient explosion.

V. Experiments

We first investigate the different mechanisms to implement spatial attention and validate the advantages of our proposed spatially-indexed mechanism over other options. Then we present the qualitative and quantitative evaluation on our model respectively, especially to validate the functionality of each module. Subsequently, we compare our model with state-of-the-art methods. Specifically, we make a statistical analysis on predicted error distributions to investigate the difference between the method based on the handcrafted features and our method with automatically learned features. Finally, we evaluate our model on disgust expression to test the generalization of our model to other facial expressions.

A. Investigation of Spatial Attention Modules

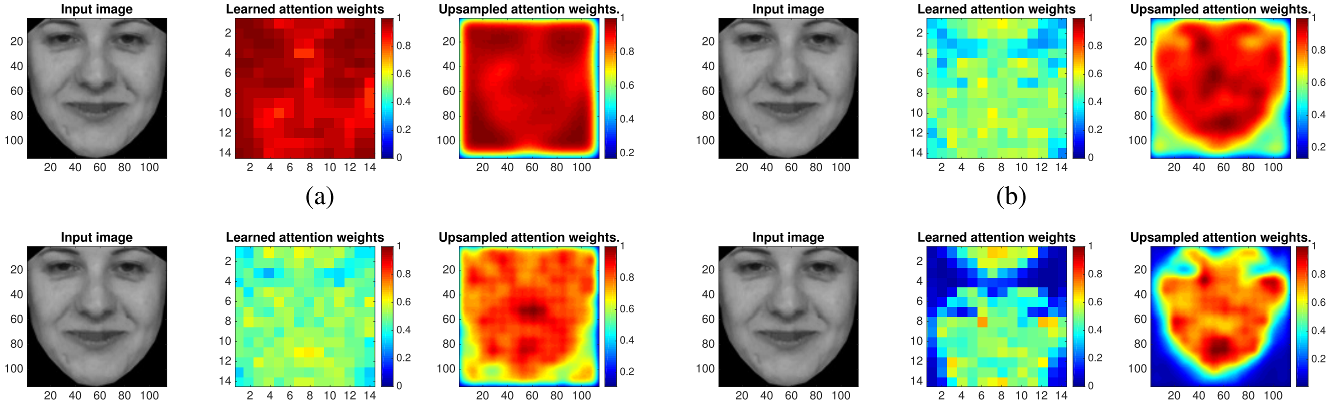
We first conduct experiments to investigate the effectiveness of the proposed spatially-indexed attention mechanism compared to other options for the spatial attentions module. Then we illustrate the effect of position where the spatial attention module is inserted in the convolutional appearance module.

1) Comparison of Different Spatial Attention Mechanisms: We propose a spatially-indexed attention mechanism indicated in Equation (1) to model the spatial attention weight $\mathbf{A}$. In order to validate the design motivation behind it, we investigate the difference of four different mechanisms:

- **Spatially-agnostic** mechanism: both $\mathbf{W}$ in the first layer and $\mathbf{u}$ in the second layer are shared across all entries of the feature map, which is the typical attention model [11].
- **Fully spatially-indexed** mechanism: both the transformation weights $\mathbf{W}$ and $\mathbf{u}$ are individually designed for each entry of the feature map.
- **Mediate spatially-indexed** mechanism: the first layer shares the transformation weights $\mathbf{W}$ while the second layer model each entry by individual weight $\mathbf{u}$.
- **Spatially-indexed** mechanism (adopted by our model): the weight $\mathbf{W}$ of the first layer is individually designed and the weight $\mathbf{u}$ in the second layer is shared.

Figure 3 presents the qualitative comparison between these four mechanisms. For each group of images, we first visualize the learned spatial attention weights for each option directly (the middle plot of each group), then we up-sample it back to the initial size of the input image by a Gaussian filter (the last plot of each group). This allow us to visualize the receptive field of attention.

It shows that the attention distribution of the spatially-agnostic mechanism appears blurred and less



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

Fig. 3. The visualization of the learned attention weights on four different spatial attention mechanisms: (a) Spatially-agnostic mechanism, (b) fully spatially-indexed mechanism, (c) mediate spatially-indexed mechanism, and (d) spatially-indexed mechanism. For each group of plots corresponding to a mechanism, we first present the original input image, subsequently we visualize the learned weights from the spatial attention module, and finally we up-sample the attention distribution back to the size of input image by a Gaussian filter. Note that the spatial attention module is inserted after the 2nd convolutional layer in this set of experiments.

TABLE I

Mean Absolute Error (Years) for Four Different Spatial Attention Mechanisms on the UvA-NEMO Smile Database

Attention Mechanism	**MAE (years)**
Spatially-agnostic	4.95 (±5.77)
Fully spatially-indexed	4.94 (±5.65)
Mediate spatially-indexed	5.01 (±5.45)
Spatially-indexed	4.74 (±5.70)

contrasting than other mechanisms. It is only able to capture the salient regions around mouth and near eyebrow. It even gives high scores to the background. It is not surprising since the receptive fields of each entry overlap, hence it is hard for the shared weights to capture the fine-grained differences. The fully spatially-indexed mechanism can roughly capture the contour of the facial regions in the images but with no highlights inside the facial area. This is because individual weights can hardly model the spatial continuity in the face. In contrast, the spatially-indexed mechanism achieves the best result among all options. Furthermore, adopting shared weights in the first layer and individual weights in the second layer (mediate spatially-indexed mechanism) is much worse than the other order. It is probably because the individual weights can hardly take effect after the smoothing by the shared weights. Therefore, our model employs the spatially-indexed mechanism, which can not only clearly distinguish the face from the background, but also capture the salient regions like the area under the eye, area around mouth and two nasolabial folds in cheeks. More examples are presented in Section V-E.1. Quantitative comparison is also provided in Table I, which demonstrates that the spatially-indexed mechanism outperforms other spatial attention mechanisms.

2) *The Effect of the Position of the Spatial Attention Module:* Theoretically, the spatial attention module can be placed after any convolutional layer in the appearance module. However, the latter convolutional layers output feature maps with larger receptive fields for each entry than the previous layers, which leads to more overlapping receptive fields of adjacent entries. As a result, each spatial attention weight also corresponds to a larger receptive field in the input image. Figure 4 shows the learned spatial attention weights after inserting the spatial attention module after different convolutional layers. In the case that the spatial attention is placed after the first layer, the distribution of learned attention weights is very noisy. This is because the small receptive fields of each attention weight results in excessively fine-grained modeling, which causes over-fitting. In contrast, placing the spatial attention module after the third convolutional layer generates more coarse and less contrasting attention weight distributions, which weakens the effect of the spatial attention module. We achieve a good balance by inserting the spatial attention module after the second convolutional layer, as shown in Figure 4.

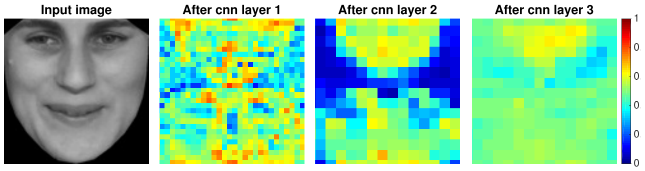


Fig. 4. The visualization of the learned spatial attention weights when placing the spatial attention module after different convolutional layers. Note that we adopt the spatially-indexed mechanism described in Section V-A.1.

3) *Investigation of Internal Scheme of Spatial Attention Module:* To investigate whether the first layer of the spatially-indexed attention module is able to capture the local detailed variation as we expect, we attempt to make a visualization on the learned weights of the first attention layer. Since the output of the first attention layer in Equation 1 is a 3D tensor with dimensions $M \times N \times d$, i.e., each pixel in the feature map of size $M \times N$ is a d-dim vector, it is hard to visualize the learned weights directly. Instead, we first perform Principal Component Analysis (PCA) on the learned attention weights,

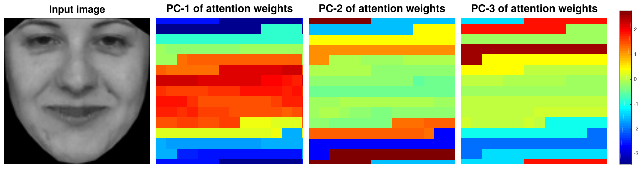

Input image
PC-1 of attention weights
PC-2 of attention weights
PC-3 of attention weights
2
1
0
-1
-2
-3

Fig. 5. The visualization of the learned attention weights of the first spatially-indexed attention layer. The Principal Component Analysis (PCA) is performed on the 256-dimension attention weights for each pixel of the attention map. The first 3 principle components (PC) which accounts for 55% variance are visualized.

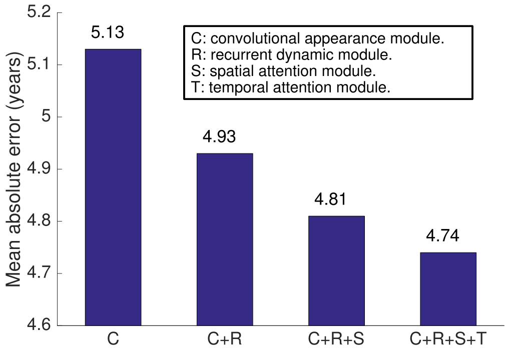


Fig. 6. Mean absolute error (years) for different functional modules on the UvA-NEMO smile database.

then we visualize the first 3 principle components which accounts for 55% the variance.

As shown in Figure 5, the first layer of the spatially-indexed attention module tends to learn the sharp variance vertically in the feature map. The whole face is roughly divided into several vertical blocks; for instance, the PC-1 of attention weights focuses more on the area around eyes and eyebrows. Resulting vertical clusters of attention scores may be explained by the horizontally-symmetric appearance of faces.

B. Quantitative Evaluation of Functionality of Each Module

Next we perform the quantitative evaluation on the four functional modules of our module, by evaluating the functionality of each module and the contribution it makes to performance of the whole architecture. To this end, we conduct ablation experiments which begin with the single convolutional appearance module in the system and then incrementally augments the system by one module at a time. When we only employ the convolutional appearance module (without modeling dynamics) in our system, we perform age estimation on each single image in the video and then average the predicted results as the final age estimation. Figure 6 presents the performance of all ablation experiments.

The individual convolutional appearance module in the system achieves an age estimation performance of 5.13 years' mean absolute error (MAE), which is an encouraging result considering the fact that it only performs appearance learning (without any dynamics information involved for age estimation). More detailed comparisons to handcrafted features are made in Table III presented in subsequent Section V-D. Equipping the system with the recurrent dynamic module results in 4.93 years' MAE, which indicates that the dynamics learning by recurrent dynamic module makes a substantial improvement. Subsequently, the spatial attention module is added into the system to capture the spatial salience in each facial image, and the MAE is decreased to 4.81 years. We will present a qualitative visualization of learned spatial attention salience in Section V-E.1 and Figure 9. Finally, including the temporal attention module, leading to our full end-to-end system, results in the best performance with 4.74 years' MAE.

It should be mentioned that the power of the temporal attention module is actually not fully exploited, since this data has been segmented to retain the smile phase mostly. Most of irrelevant segments before and after the smile have been removed. The temporal attention module will be shown qualitatively to be capable of capturing the key smile segment precisely in Section V-E.2 and Figure 10. Hence more improvement by the temporal attention module is expectable given temporally noisier data.

TABLE II

MEAN ABSOLUTE ERROR (YEARS) FOR DIFFERENT TEMPORAL PHASES FOR AGE ESTIMATION ON THE UVA-NEMO SMILE DATABASE

Temporal Phase(s)	**MAE (years)**
Onset	8.41 (±8.95)
Apex	5.91 (±7.05)
Offset	8.75 (±9.68)
Onset + Apex	5.04 (±5.80)
Apex + Offset	5.13 (±5.83)
Onset + Apex + Offset	4.74 (±5.70)

C. Effect of Temporal Phases

A facial expression consists of three non-overlapping temporal phases, namely: 1) the onset (neutral to expressive), 2) apex, and 3) offset (expressive to neutral), respectively. To understand the informativeness of different temporal phases in age estimation, we evaluate the accuracy on each temporal phase using our model pre-trained on the whole expression (from the beginning of onset to the end of offset).

As shown in Table II, the apex phase contributes most to the task of age estimation. It is reasonable since smiling faces tend to contain more discriminative features like wrinkles than other phases. It is also consistent with the distribution of temporal attention weights showed in Section V-E.2 and Figure 10. As expected, the combined use of consecutive phases leads to a better performance since these phases have different temporal dynamics and appearance patterns that can provide additional information. Therefore, we may claim that our approach can still be used (with a relative decrease in accuracy) even if the whole duration of an expression is not captured.

D. Comparison to Other Methods

Next, the model is compared with existing methods for age estimation, that can be applied to sequential images (videos). According to the mechanism of the feature design, these baseline methods can be classified into four categories as listed in Table III:

TABLE III
MEAN ABSOLUTE ERROR (YEARS) FOR DIFFERENT METHODS ON THE UvA-NEMO SMILE DATABASE

Method		**MAE (years) for Different Age Ranges**								
		0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	All
Spatio-temporal	VLBP [62]	10.69	12.95	15.99	18.54	18.43	16.58	23.80	26.59	15.70 (±12.40)
	LBP-TOP [63]	9.71	11.01	14.19	15.88	16.75	15.29	19.70	23.71	13.83 (±10.97)
Dynamics	Deformation [10]	4.85	8.72	12.22	13.06	13.53	11.55	14.13	17.82	10.81 (±8.85)
	Displacement [9]	5.42	9.67	11.98	14.53	12.77	15.42	20.57	20.35	11.54 (±11.49)
Appearance	IEF, Neutral [10]	3.27	3.96	4.86	6.11	5.53	8.24	12.89	13.35	5.21 (±4.48)
	IEF, Fusion [10]	3.54	4.38	5.43	6.74	6.01	8.96	13.52	14.05	5.71 (±4.65)
	LBP, Neutral [10]	3.86	4.65	5.87	6.82	5.58	8.32	9.84	11.9	5.67 (±4.97)
	LBP, Fusion [10]	4.18	4.99	6.31	7.37	6.19	8.67	10.34	12.93	6.12 (±5.11)
	CNNs	2.30	3.09	4.69	5.26	5.82	10.58	17.20	23.29	5.13 (±5.68)
Appearance+Dynamics	IEF+Dynamics [10]	3.96	4.45	4.50	5.29	4.74	6.85	12.43	11.94	5.00 (±4.25)
	LBP+Dynamics [10]	3.49	4.68	5.13	5.85	5.24	7.05	12.17	12.00	5.29 (±4.36)
	SIAM (Our model)	1.79	2.45	4.26	4.97	5.36	11.86	16.43	23.12	**4.74** (±5.70)
	Number of Samples	158	333	215	171	250	66	30	17	1240

- **Spatio-temporal**: Hadid et al. [43] propose to employ the spatio-temporal information for classifying age intervals. Particularly, they extract volume LBP (VLBP) features and feed them to a tree of four SVM classifiers. Another method using spatio-temporal information is proposed to extract the LBP histograms from Three Orthogonal Planes (LBP-TOP): XY (two spatial dimensions in a single image), XT (X dimension in the image and temporal space T) and YT [61]. Thus the information in temporal domain is utilized together with information in spatial image domain. These two spatio-temporal methods are implemented for age estimation as baselines by Dibeklioğlu *et al.* [10], from where we report the results.
- **Appearance + Dynamics**: Dibeklioğlu *et al.* [10] is the first study which leverages both the facial dynamics and appearance information for age estimation. They propose several handcrafted dynamics features specifically for facial expressions and combine them with appearance features to perform age estimation through a hierarchical architecture. They combine their dynamics features with four different kinds of appearance descriptors in their system. Among them we select two combinations with the best performance as our baselines: dynamics + IEF (Intensity-based encoded aging features [24]) and dynamics + LBP (local binary patterns) [16].
- **Dynamics**: We incorporate the baseline models using sole dynamics information. Following Dibeklioğlu *et al.* [10], we compare the deformation-based features and the displacement dynamics features [9].
- **Appearance**: We also compare our method to appearance-based approaches that solely employ IEF and LBP features [10], where (1) the neutral version estimates the age on the first frame of a smile onset (neutral face); (2) the fusion version averages age estimations from the first and the last frame of a smile onset (a neutral and an expressive face, respectively) for the final prediction. Furthermore, we evaluate a modified version of our method that uses only convolutional appearance module.

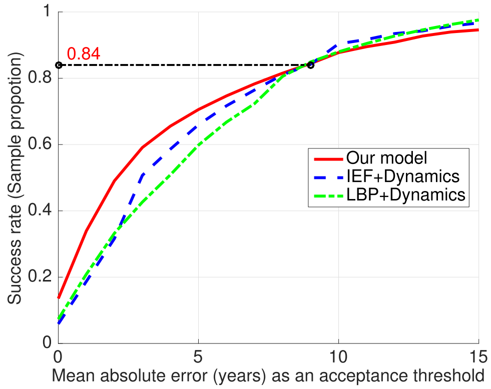


Fig. 7. Cumulative distribution of the mean absolute error for different models using both appearance and dynamics on the UvA-NEMO Smile Database.

1) Performance Comparison: Table III shows the mean absolute errors (MAE) obtained by four categories of age estimation methods mentioned before. Our model achieves the best performance considering all the age ranges with the minimum MAE of 4.74 years. While spatio-temporal methods perform worst, the methods utilizing both appearance and dynamics are more accurate than the methods based on sole appearance or dynamics. It illustrates the importance of both appearance and dynamics to the task of age estimation.

In particular, the performance of our model is better than the other two methods using both appearance and dynamics: *IEF+Dynamics* and *LBP+Dynamics* [10]. Figure 7 shows the success rate as a function of the MAE for these three methods. For age deviations up to nine years, our model outperforms the other two methods. For larger age deviations, the model is slightly worse. Our model suffers from some severe estimation

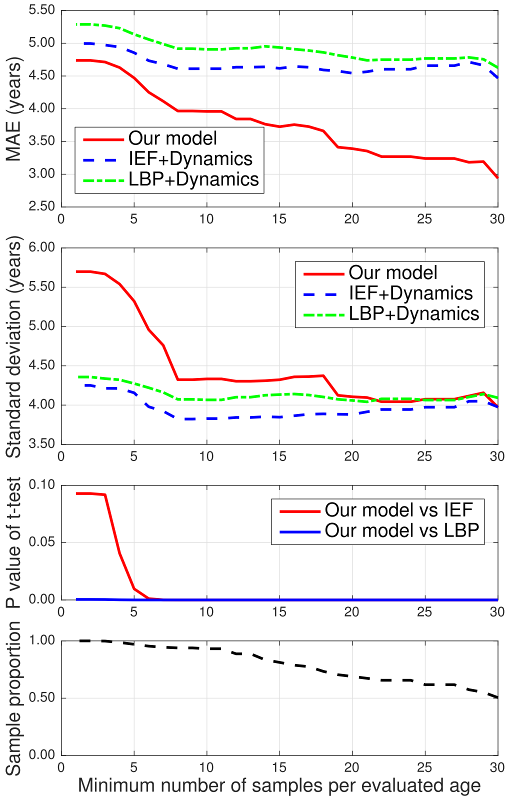


Fig. 8. Performances of three methods using both the appearance and dynamics on the UvA-NEMO Smile Database as a function of minimum number of samples per evaluated age.

errors on a few samples. These samples appear to be from the high-age groups, where our model is severely hampered by the low sample size in these groups. Figure 2 and Table III both show that the number of samples in these age ranges is much less than that in the younger age ranges. Compared to handcrafted features used in *IEF+Dynamics* and *LBP+Dynamics* [10], the convolutional appearance module and recurrent dynamic module in our model require more training data.

2) The Effect of Training Size per Age: To investigate the effect of training size per age on the performance of three *Appearance+Dynamics* models, we conduct experiments where the data are reduced by removing ages for which a low number of samples are available. The results are presented in Figure 8. The experiments begin with the case that all the samples involved (threshold = 1). As the threshold (minimum number of samples) is increased, the performance gap between our model and the other methods becomes larger: the MAE of our model decreases much faster than the other two methods and the variance also drops deeply at the beginning of the curve. A t-test shows that our model significantly outperforms other methods when the threshold on the number of sample is larger than 5 ($p < 0.01$). They are actually quite encouraging results, since these results in turn indicate that larger data tend to explore more potential of our model and make it more promising than the other two methods on the task of age estimation.

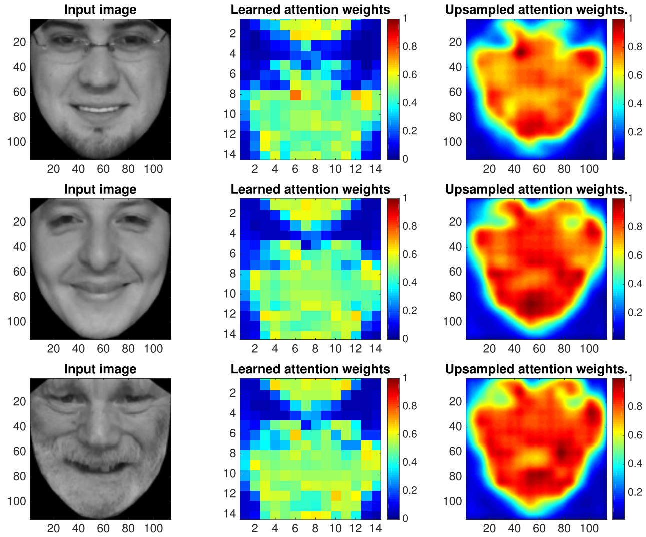


Fig. 9. The heat map visualization of the learned attention weights by our spatial attention module. For each subject, the middle plot corresponds to attention weights and the last plot is the up-sampled attention weight distribution back to the size of initial input image by a Gaussian filter.

E. Qualitative Evaluation of Attention Modules

In this section, we qualitatively evaluate the spatial attention module and temporal attention module by visualizing the learned attention weights.

1) Spatial Attention Module: In Figure 9, three sample images are presented to visualize the results of spatial attention module. For this module the optimal configuration is used: it uses the spatially-indexed mechanism and includes the spatial attention module after the second convolutional layer. The images on the left are the original inputs, the images in the middle are heat maps of the attention values, and the images on the right are upscaled heat maps to the original resolution. These heat images show that our spatial attention module is able to not only discriminate the core facial region from the background accurately, but also capture the salient parts inside the facial region. Specifically, the area under eyes, nasal bridge, two nasolabial folds, and area around mouth (especially mentolabial sulcus) are detected as the salient parts. It is reasonable since these areas tend to generate wrinkles easily when smiling, which are discriminative features for the task of age estimation.

2) The Temporal Attention Module: To demonstrate the temporal salience detection, the learned temporal attention weights for several representative frames from two video samples are shown in Figure 10. Each row shows the smile progression from a neutral state to smiling and back to neutral. For the neutral faces at the beginning, our module predicts very small temporal attention weights. As the degree of smiling increases, the attention weight goes up accordingly, until to the peak value. It should be noted that the attention value grows rapidly with the appearance of the two nasolabial folds, which is consistent with the facial salience captured by the spatial attention module (shown in Figure 10). Then the attention value decreases with the recession of smiling.

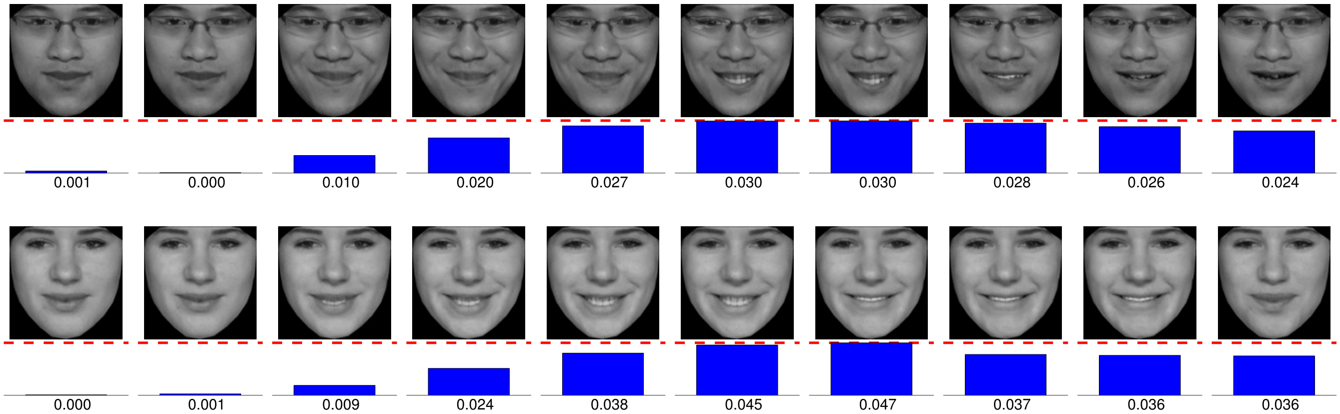


Fig. 10. The visualization of learned temporal attention weights for examples from test set of UvA-NEMO smile database. Higher bar indicates larger attention weight. The attention weight is indicated for representative frames. Our temporal attention modules is able to cut off neutral faces in the beginning phase of the smiling and assign higher value to frames with higher degree of smiling. Note that attention weights for all frames in a video are normalized (Equation 4) to make them sum up to 1. The red dash line indicates the maximum value of attention weights.

TABLE IV

Mean Absolute Error (Years) for Different Methods on the UvA-NEMO Disgust Database and the Posed UvA-NEMO Smile Database

Database	Method	MAE (years)
Posed Disgust	IEF+Dynamics [10]	5.06 (±4.45)
	LBP+Dynamics [10]	5.19 (±4.51)
	SIAM (Our model)	6.96 (±7.24)
Posed Smile	IEF+Dynamics [10]	4.91 (±4.17)
	LBP+Dynamics [10]	5.16 (±4.30)
	SIAM (Our model)	6.41 (±6.65)

However, the value still retains a relatively high value at the last frame. It is partially because the hidden representation of the last frames contains the information of all previous frames as well as key frames about smiling, hence they are still helpful for age estimation. Besides, the smile videos in the given database do not end with a perfectly netural face. Otherwise, the attention weight would continuously decrease for the latter neutral faces.

F. Application to Disgust Expression

To evaluate whether the proposed approach can be generalized to other facial expressions, we conduct experiments on the UvA-NEMO Disgust Database, which is composed of posed disgust expressions. Please note that the UvA-NEMO Smile and Disgust databases are the only facial expression video databases for age estimation.

Table IV presents the mean absolute errors (MAE) by our model as well as the state-of-the-art methods that handcraft several facial dynamics features and combine them with IEF [24] and LBP [16] appearance features. Unexpectedly, our model performs worse than the state-of-the-art methods, which is in contrast to the experimental results on the UvA-NEMO smile database shown in Table III. We consider two underlying reasons leading to poor performance of our model on disgust database. First, it is due to the relatively small data size of disgust database. The disgust database consists of 518 videos, which is only around half size of the smile database but shares the same age interval. It is revealed in Section V-D.2 that the training size per age significantly affects the performance of our model, which in turn explains that the poor performance of our model on the disgust database may be caused by the limited training data size. Besides, it is important to note that the carefully designed handcrafted features used in the competitor methods do not require large-scale training data since they rely on priori knowledge from the literature. Thus, we may claim that it is not fair to directly compare handcrafted features with deep models using relatively small training data.

Secondly, the disgust database has only posed expressions while the smile database consists of both spontaneous and posed expressions. One may speculate that the posed expressions are not as discriminative as the spontaneous ones for age estimation. However, [9] reports a better age estimation performance for posed expressions. Therefore, to further assess the negative effect of small data size by discarding the probable influence of expression spontaneity, we split the UvA-NEMO Smile Database into the spontaneous and posed smile sets, and evaluate our model solely using the posed set (643 posed smiles). Table IV shows that our model performs much worse on the posed smiles (MAE $= 6.41$ years) in comparison to the whole (spontaneous + posed) smile dataset (MAE $= 4.74$ years). This result is consistent with our conjecture. Hence, we can claim that the performance our model achieves on the UvA-NEMO Disgust Database is reasonable due to the small data size, and a significant accuracy improvement can be achieved using a larger training set.

VI. Conclusion

In this work, we present an attended end-to-end model for age estimation from facial expression videos. The model employs convolutional networks to learn the effective appearance features and feed them into recurrent networks to learn the temporal dynamics. Furthermore, both a spatial attention mechanism and a temporal attention mechanism are added to the model. The spatial attention can be integrated seamlessly into the convolutional layers to capture the salient facial regions in each single image, while the temporal attention is incorporated in recurrent networks to capture the salient temporal frames. The whole model can be trained readily in an end-to-end manner. Provided that a sufficient number of samples are available for training, we show the strong performance of our model on a large smile database. Specifically, our model makes a substantial improvement over the state-of-the-art methods. Furthermore, we assess the applicability of the proposed method for disgust expression.

In future work, we aim to leverage the pre-trained convolutional neural networks on large image data for the appearance learning instead of training our convolutional appearance module from scratch. This would not only accelerate the training speed but also allows employing quite deeper architectures and abundant existing image data to improve the performance of the appearance learning.

References

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

[2] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 621–628, Feb. 2004.

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

[4] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.

[5] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 256–263.

[6] A. C. Gallagher and T. Chen, "Estimating age, gender, and identity using first name priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

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

[8] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe, "Apparent and real age estimation in still images with deep residual regressors on Appa-real database," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May/Jun. 2017, pp. 87–94.

[9] H. Dibeklioğlu, T. Gevers, A. A. Salah, and R. Valenti, "A smile can reveal your age: Enabling facial dynamics in age estimation," in *Proc. 20th Int. Conf. Multimedia*, 2012, pp. 209–218.

[10] H. Dibeklioğlu, F. Alnajar, A. A. Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1928–1943, Jun. 2015.

[11] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, vol. 37, Jul. 2015, pp. 2048–2057.

[12] W. Pei, T. Baltrusaitis, D. M. Tax, and L.-P. Morency, "Temporal attention-gated model for robust sequence classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6730–6739.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[14] T. R. Alley, *Social and Applied Aspects of Perceiving Faces*. London, U.K.: Psychology Press, 2013.

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

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

[17] J. Ylioinas, A. Hadid, X. Hong, and M. Pietikäinen, "Age estimation using local binary pattern Kernel density estimate," in *Proc. Int. Conf. Image Anal. Process.*, 2013, pp. 141–150.

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

[19] P. Thukral, K. Mitra, and R. Chellappa, "A hierarchical approach for human age estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1529–1532.

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

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

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

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

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

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

[26] C. Li, Q. Liu, J. Liu, and H. Lu, "Ordinal distance metric learning for image ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1551–1559, Jul. 2015.

[27] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 534–541.

[28] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2015, pp. 34–42.

[29] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah, "Kernel ELM and CNN based facial age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2016, pp. 80–86.

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, vol. 1, no. 3, p. 6.

[31] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 144–157, 2016.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[33] E. Agustsson, R. Timofte, and L. Van Gool, "Anchored regression networks applied to age estimation and super resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1643–1652.

[34] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Pattern Recognit.*, vol. 66, pp. 106–116, Jun. 2017.

[35] X. Wang and C. Kambhamettu, "Age estimation via unsupervised neural networks," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–6.

[36] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen, "Automatic age estimation from face images via deep ranking," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 55.1–55.11.

[37] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4920–4928.

[38] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5183–5192.

[39] G. Guo and X. Wang, "A study on human age estimation under facial expression changes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2547–2553.

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

[41] G. Guo, R. Guo, and X. Li, "Facial expression recognition influenced by human aging," *IEEE Trans. Affect. Comput.*, vol. 4, no. 3, pp. 291–298, Jul. 2013.

[42] Z. Lou, F. Alnajar, J. M. Alvarez, N. Hu, and T. Gevers, "Expression-invariant age estimation using structured learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 365–375, Feb. 2017.

[43] A. Hadid, "Analyzing facial behavioral features from videos," in *Proc. Int. Workshop Hum. Behav. Understand.*, 2011, pp. 52–61.

[44] N. Ramanathan, R. Chellappa, and S. Biswas, "Computational methods for modeling facial aging: A survey," *J. Vis. Lang. Comput.*, vol. 20, no. 3, pp. 131–144, 2009.

[45] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the FG-NET ageing database," *IET Biometrics*, vol. 5, no. 2, pp. 37–46, May 2016.

[46] L. Yao *et al.*, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4507–4515.

[47] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[48] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, 2015, pp. 2017–2025.

[49] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1841–1850.

[50] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3391–3399.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[52] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[53] P. J. Werbos, "Generalization of backpropagation with application to a recurrent gas market model," *Neural Netw.*, vol. 1, no. 4, pp. 339–356, 1988.

[54] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at Me? spontaneous versus posed enjoyment smiles," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 525–538.

[55] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[56] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, Jun. 2013, pp. 354–361.

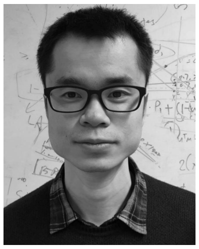
[57] P. F. Velleman, "Definition and comparison of robust nonlinear data smoothing algorithms," *J. Amer. Statist. Assoc.*, vol. 75, no. 371, pp. 609–615, 1980.

[58] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*. [Online]. Available: https://arxiv.org/abs/1207.0580

[59] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[60] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, "Advances in optimizing recurrent networks," in *Proc. ICASSP*, May 2013, pp. 8624–8628.

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



Wenjie Pei received the Ph.D. degree from the Delft University of Technology, The Netherlands, in 2018. He is currently an Assistant Professor with the Harbin Institute of Technology, Shenzhen, China. Before joining the Harbin Institute of Technology, he was a Senior Researcher of computer vision with Tencent Youtu X-Lab. In 2016, he was a Visiting Scholar with the Carnegie Mellon University. His research interests include computer vision and pattern recognition including sequence modeling, deep learning, and video/image captioning.

Hamdi Dibeklioğlu (S'08–M'15) received the Ph.D. degree from the University of Amsterdam, Amsterdam, The Netherlands, in 2014. He is currently an Assistant Professor with the Computer Engineering Department, Bilkent University, Ankara, Turkey, as well as being a Research Affiliate with the Pattern Recognition & Bioinformatics Group of Delft University of Technology, Delft, The Netherlands. Before joining Bilkent University, he was a Postdoctoral Researcher with the Delft University of Technology. His research focuses on computer vision, pattern recognition, affective computing, and computer analysis of human behavior. He is also a Program Committee Member for several top tier conferences in these areas. He was the Co-chair of the Netherlands Conference on Computer Vision in 2015, the Local Arrangements Co-chair of the European Conference on Computer Vision in 2016, the Publication Co-chair of the European Conference on Computer Vision in 2018, and the Co-chair of the eNTERFACE Workshop on Multimodal Interfaces in 2019.

Tadas Baltrušaitis received the bachelor's and Ph.D. degrees in computer science. He is a Senior Scientist with Microsoft Corporation. Before joining Microsoft, he was a Postdoctoral Associate with Carnegie Mellon University. His Ph.D. research focused on automatic facial expression analysis in especially difficult real-world settings. His primary research interests include the automatic understanding of non-verbal human behaviour, computer vision, and multimodal machine learning.

David M. J. Tax studied physics at the University of Nijmegen, The Netherlands, in 1996, and received master's degree with the thesis "learning of structure by Many-take-all Neural Networks". After that he had the Ph.D. degree at the Delft University of Technology in the Pattern Recognition Group, under the supervision of R.P.W. Duin. In 2001, he promoted with the thesis one-class classification. After working for two years as a Marie Curie Fellow with the Intelligent Data Analysis Group, Berlin, he is currently an Assistant Professor with the Pattern Recognition and Bioinformatics Group, Delft University of Technology. He is also lecturing in courses pattern recognition, machine learning, probability and statistics, and stochastic processes. His main research interest is in the learning and development of detection algorithms and (one-class) classifiers that optimize alternative performance criteria like ordering criteria using the area under the ROC curve or a Precision-Recall graph.