Personalized Modeling of Facial Action Unit Intensity

Shuang Yang[1], Ognjen Rudovic[1], Vladimir Pavlovic[2], and Maja Pantic[1,3]

[1] Comp. Dept., Imperial College London, UK
[2] Dept. of Computer Science, Rutgers University, USA
[3] EEMCS, University of Twente, The Netherlands

Abstract. Facial expressions depend greatly on facial morphology and expressiveness of the observed person. Recent studies have shown great improvement of the personalized over non-personalized models in variety of facial expression related tasks, such as face and emotion recognition. However, in the context of facial action unit (AU) intensity estimation, personalized modeling has been scarcely investigated. In this paper, we propose a two-step approach for personalized modeling of facial AU intensity from spontaneously displayed facial expressions. In the first step, we perform facial feature decomposition using the proposed matrix decomposition algorithm that separates the person's identity from facial expression. These two are then jointly modeled using the framework of Conditional Ordinal Random Fields, resulting in a personalized model for intensity estimation of AUs. Our experimental results show that the proposed personalized model largely outperforms non-personalized models for intensity estimation of AUs.

1 Introduction

Facial expressions communicate emotions, clarify and stress what is being said, and signal comprehension, disagreement and intentions. Machine understanding of facial expressions could revolutionize user interfaces for artifacts such as robots, mobile devices, cars, and conversational agents [1]. The Facial Action Coding System (FACS) [2] is the most comprehensive, anatomically-based system for encoding expression. FACS defines 33 atomic facial muscle actions named Action Units (AUs), where the intensity of each AU ranges from being absent to having maximal intensity on a six-point ordinal scale. The AU intensity coding is usually carried out by trained human FACS coders. Nevertheless, this process is tedious and error prone [3]. This is also true because the facial AU intensity reflects large variability in persons' facial morphology and their expressiveness, head-movements, illumination changes, and, to some extent, the coders' bias. All this makes the automated estimation of the AU intensity highly challenging.

In this paper, we investigate the influence of individual differences on automated intensity estimation of facial AUs. The changes in the intensity of AUs are reflected in subtle variability in the person's facial appearance. This can differ significantly among persons mainly because the muscular contractions of the face are combined with their individual physical characteristics [2]. Also, each person may have a different level of expressiveness (e.g., extrovert vs. introvert). For example, persons gesticulate differently and while for some the appearance of cheek dimples is their most intense smile,

G. Bebis et al. (Eds.): ISVC 2014, Part II, LNCS 8888, pp. 269–281, 2014.

for others that is just the slight intensity smile. This, in turn, makes it difficult to grasp what constitutes the maximal level of appearance change for each person.

Most of existing works on automated analysis of facial expressions employ generic (non-personalized) classifiers that do not explicitly account for individual differences (e.g. [4,5,6,7]). These works are based on different facial features, obtained by attempting to remove the person's identity information. This is typically done by applying decomposable models to extract the expression specific variation from images [4,5], or by subtracting the first frame in the image sequence from the remaining frames [6,7]. Then, different generic classifiers are applied to such 'un-personalized' facial features. However, separating the identity from expression variation is not trivial, and often results in a loss of expression-related information. This, in turn, adversely affects the models' performance in the expression recognition tasks. To tackle this, various methods have been proposed for personalized modeling of facial expressions.

The personalized models can be divided into three groups: the person-dependent models ([8,9,10]), the person-adaptable models ([11,12,13,14]), and the models that use personalized facial features ([15]). The first group uses data of both the training and test persons during learning, and these models are typically tailored to each person (training and test). For instance, [8] proposed person-dependent models, based on template-matching classifiers, for recognizing the facial expressions of six basic emotions and neutral. Similarly, [9] used person-specific facial expression data to adapt the Support Vector Machine (SVM) classifier to each individual, with the aim of predicting topical relevance in the context of information search and retrieval. [10] proposed a person-dependent graph-fitting method for facial feature tracking, the output of which was used to derive person-dependent facial features for emotion recognition, based on the matching of the personalized facial action graphs. All these approaches achieved better performance in target tasks than generic models; however, separate models needed to be trained for each person.

The second group is based on transfer learning/multi-task techniques, where the model parameters are first learned using the expression data of training persons, and then adapted to the test person during inference. For instance, [11] proposed an adaptable non-linear model, based on functional analysis, for recognizing facial expressions of six basic emotions and neutral. The authors first trained a generic classifier, and then during inference used a small set of images of the test person displaying seven emotional states to retrain the model parameters. In [12], the authors proposed transfer learning models that capture commonalities across a set of training persons and also learn the way each individual instantiates these commonalities. These models are based on optimization problems that use regularizers on the task parameters, encoding the relationships among the tasks (i.e., persons). By comparing the models on the pain expression recognition task, the proposed adaptable models outperformed generic models based on the subtractive method. Similarly, [13] proposed two transfer learning algorithms: inductive and transductive transfer learning for detection of pain and AUs, in both a semi-supervised and unsupervised settings. [16] proposed a personalized model for AU detection that is based on the Kernel Mean Matching technique. In this approach, an iterative minimization procedure is proposed to adapt the hyperplanes of generic SVM, trained using the labeled source data, to the target person. [14] proposed another approach for personalizing models by first decoupling different factors (i.e., emotional states and identity) using multi-task learning, under assumption that the emotion related tasks are orthogonal to the identity tasks. Then, the identity tasks are used as *auxiliary* tasks to improve learning of the *principal* tasks, i.e., emotion expressions.

The representative of the third group is the recently proposed Context-sensitive Conditional Ordinal Random Field (cs-CORF) [15], where the authors modeled the person identity as a context factor affecting the estimation of intensity of AUs. Specifically, the authors defined the context covariate effects that encode the person's characteristics (extracted from the first neutral frame in an image sequence of the varying facial AU intensity), and context-free covariate effects (as those used in generic models and obtained by subtracting the first frame in the sequence from the rest). These were then jointly modeled in the cs-CORF model, outperforming generic CORF and other generic classification models [3,17,18]. However, a downside of this approach is that the identity features are derived from each sequence separately, thus, not capturing the commonalities of target person among multiple sequences of the same person. Also, decomposition of the identity and expression is performed in the common subspace obtained by PCA. Yet, these two may better be represented by different subspaces, as they represent different types of variation in the data.

To address some of the limitations mentioned above, we propose a latent factor analysis model for personalized estimation of the AU intensity. Specifically, we propose an iterative matrix decomposition algorithm for learning the identity and AU-intensity-specific latent factors which jointly generate the observed faces. The identity factor represents the between-person variance in our data, and is constant for each person in the dataset. On the other hand, the AU-intensity-related factor varies across sequences of target persons, thus, accounting for the within-person variation due to the AU intensity changes. Once learned, these two latent factors are jointly modeled using the CORF framework for dynamic estimation of facial AU intensity [15]. The proposed decomposition algorithm can easily handle a large number of image sequences, with the fast convergence rate. We show in our experiments that by personalizing the CORF model [7] via inclusion of the identity latent factors, we achieve better estimation of AU intensity compared to generic CORF model that ignores the identity of target persons, as well as of personalized CORF model that performs the feature decomposition in the common subspace (i.e., the cs-CORF model [15]).

The remainder of the paper is organized as follows. Sec.2 describes the proposed approach to personalized modeling of the facial AU intensity. Sec.3 shows the results of the experimental evaluation, and Sec.4 concludes the paper.

2 Methodology

In this section, we propose our approach to personalized modeling of intensity of facial AUs from image sequences. We first propose a matrix decomposition algorithm for separating the measurement features into two latent factors: the person's identity and facial expression. Then, we describe how these factors can be modeled within the CORF framework, resulting in a personalized model for structured prediction of facial AU intensity. In what follows, we assume we are given n image sequences $\mathcal{D} = \{(\mathbf{x}^l, \mathbf{y}^l)\}_{l=1}^{n}$, where the observations, denoted by $\mathbf{x}$, serve as input covariates for predicting $\mathbf{y}$. Furthermore, each observation in an image sequence is denoted as x_{ijk}, where $i = 1 \ldots N_s$ is the person index and N_s is the number of persons in the dataset, $j = 1 \ldots N_i$ is the number of sequences of the i-th person, and $k = 1 \ldots T_j$ is the time index for the observations in the j-th sequence of the i-th person. The output variable y is structured in the same way, and at each time step is assumed to take one of R different (ordinal) categories, i.e., the AU intensity levels, as $y_{ijk} \in \{1, \ldots, R\}$.

2.1 Personalized Facial Feature Decomposition

Our goal is to decompose facial features extracted from image sequences of different persons with varying facial AU intensity levels into person-specific characteristics, which are assumed to be constant across sequences of target person, and AU-intensity-specific characteristics, which are assumed to vary across image sequences of that person. In other words, we seek to find two low-dimensional spaces: the identity space and the AU-intensity-specific space, which, together, generate the observed facial features. Various approaches that address this task have been proposed [19,4,20,21]. These are based on the tensor representation of different factors (i.e., identity, pose, illumination, and expression), decoupling of which is attained by means of multilinear generalizations of Singular Value Decomposition (SVD). These approaches subsume as special cases the simple linear (1-factor) analysis such as principal components analysis (PCA), as well as bilinear (2-factor) analysis [5]. In particular, a representation named *TensorFaces* has firstly been introduced for learning multilinear models of facial image ensembles resulting from interaction of any number of underlying factors (e.g., pose and expression). This multilinear representation yields superior facial recognition rates compared to the standard linear approaches (e.g., PCA/eigenfaces [22]). Although such approaches may seem a suitable choice for our task, decomposition of tensors into different factors becomes inefficient or even intractable when dealing with high-dimensional input features, and, in particular, with image sequences.

Instead of using tensors, some authors employ the additive factor analysis approach, where the observed data is represented as a linear sum of different latent factors. For instance, [5] proposed a two-factor model for decoupling the identity and facial expression for 3D facial expression recognition. For the task of face recognition, [23] proposed the probabilistic Linear Discriminant Analysis (pLDA) model, a latent factor analysis model for learning identity latent variables and those not related to the identity. The main assumption behind this approach is that the observed faces can be generated from the latent identity variables by a noisy process. The learning of the latent variables (the identity and other identity-unrelated factors such as expression, illumination, etc.) is performed via an Expectation-Maximization algorithm. Again, although this algorithm is computationally efficient for the face recognition tasks, as there are usually only a few examples of target persons available during training, it becomes computationally intractable when dealing with large number of data, as in the case of image sequences.

To address the limitations mentioned above, we propose a latent factor model for personalized facial feature decomposition that can deal efficiently with image sequences.

Formally, we decompose the observed facial features within an image sequence into the person-specific characteristics (identity) and facial AU intensity characteristic as:

$$\mathbf{x}_{ijk} = \mu + \mathbf{F}h_i + \mathbf{G}w_{ijk}, \tag{1}$$

where we assumed the noise-free case. We also set the mean of the data, μ, to zero. $\mathbf{F}$ is a factor matrix with the basis vectors of the between individual subspace in its columns, and h_i is the latent identity variable that is constant across all image sequences $j = 1, \ldots, N_i$ of person i. Likewise, the matrix $\mathbf{G}$ contains a basis for the within-individual subspace, and w_{ijk} is the latent variable accounting for variation due to changes in the AU intensity levels at each time step k within a sequence. Both variables h and w are assumed to live in a low-dimensional space of size D_h and D_w, defined by the projection matrices $\mathbf{F}$ and $\mathbf{G}$, respectively.

To perform the decomposition in (1), we need to estimate ($\mathbf{F}$, $\mathbf{G}$), and for each x_{ijk} find the corresponding values of the latent variables (h_i, w_{ijk}). To this end, we propose an iterative algorithm, which is described in Alg.1. We briefly explain the algorithm. In the first step, we compute the data clusters, where the centroid of each cluster m_i, $i = 1 \ldots N_s$, represents the average of the facial features for person i in the dataset. This is followed by the eigen decomposition (*eig*) of the covariance matrix computed from the mean-normalized centroids of each person. This gives us a closed-form solution for the person identity matrix $\mathbf{F}$, computed using the obtained eigenvectors U_F and eigenvalues, stored in the diagonal matrix S_F. The latent identity factors for each person are then determined by projecting the person centroids onto the first D_h eigenvectors from $\mathbf{F}$. Once we estimated the identity component, we next seek to find the latent factor responsible for changes in the AU intensity levels. Thus, in the second step, we estimate $(\mathbf{G}, w_{ijk})$ in a similar way, but this time by applying *eig* to the covariance matrix of the data residuals obtained after subtracting the identity component ($F_{1:D_h} \cdot h_i$) from the observed facial features. By projecting the AU intensity residuals onto the first D_w eigenvectors in $\mathbf{G}$, we obtain intensity-related latent factors w_{ijk}. Finally, we compute the identity residuals by subtracting the AU intensity component from the observed facial features, and then apply Step 1 to these residuals. We alternate between Steps 1 and 2 until convergence of the reconstruction error. Intuitively, the algorithm searches for a decomposition that jointly best explains the observed facial features, while trying to maximally separate the between-person variance (the identity component) from within-person variance (the AU intensity component). For the data that we used in this paper, the algorithm typically converges in less than 150 iterations.

2.2 Personalized Conditional Ordinal Random Fields (p-CORF)

In this section, we personalize the CORF [7,15] model for dynamic estimation of facial AU intensity levels. The CORF model is adaptation of the linear-chain CRF [24] model attained by setting CRF's node features using the modeling framework of ordinal regression [25]. In this way, the monotonicity constraints are imposed on the ordinal labels (i.e., intensity levels of AUs). Formally, given the j-th image sequence of person i, $\mathbf{x}_{ij} = \{x_{ij1}, \ldots, x_{ijT_j}\}$, and the corresponding AU intensity labels, $\mathbf{y}_{ij} = \{y_{ij1}, \ldots, y_{ijT_j}\}$, we write the conditional distribution $P(\mathbf{y}_{ij}|\mathbf{x}_{ij})$ of the CORF model

Algorithm 1. Personalized Facial Feature Decomposition

```
Learning
Inputs: $x_{ijk}, i = 1...N_s, j = 1...N_i, k = 1...T_j, D_h, D_w$.
Initialize: $\forall ijk : \Delta x_{ijk} = x_{ijk}$.
repeat
    Step 1: Compute $(\mathbf{F}, h)$
        $\forall i : m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{T_j} \sum_{k=1}^{T_j} \Delta x_{ijk}, \hat{m} = \frac{1}{N_s} \sum_{i=1}^{N_s} m_i, M = [m_1 - \hat{m}, ..., m_{N_s} - \hat{m}]$
        $[U_F, S_F] = eig(M^T M), \mathbf{F} = M U_F S_F^{-1}, \forall i : h_i = \mathbf{F}_{1:D_h}^T (m_i - \hat{m})$
    Step 2: Compute $(\mathbf{G}, w)$
        $\forall ijk : q_{ijk} = \Delta x_{ijk} - \mathbf{F}_{1:D_h} h_i - \hat{m}, \hat{q} = \frac{1}{N} \sum_{ijk} q_{ijk}, Q = [q_{111} - \hat{q}, ..., q_{N_s N_{N_s} T_{N_{N_s}}} - \hat{q}]$
        $[\mathbf{G}, S_G] = eig(QQ^T), \forall ijk : w_{ijk} = \mathbf{G}_{1:D_w}^T (q_{ijk} - \hat{q}), \Delta x_{ijk} = x_{ijk} - \mathbf{G}_{1:D_w} w_{ijk} - \hat{q},$
until $err = \frac{1}{N} \sum_{ijk} ||x_{ijk} - \mathbf{F}_{1:D_h} h_i - \hat{m} - \mathbf{G}_{1:D_w} w_{ijk} - \hat{q}||^2$ converges.
Outputs: $\mathbf{F}, \mathbf{G}, \forall ijk : h_i, w_{ijk}$.
```

as the Gibbs form clamped on the observations $\mathbf{x}_{ij}$:

$$P(\mathbf{y}_{ij}|\mathbf{x}_{ij}, \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_{ij}; \boldsymbol{\theta})} e^{s(\mathbf{x}_{ij}, \mathbf{y}_{ij}; \boldsymbol{\theta})}, \tag{2}$$

where $Z(\mathbf{x}_{ij}; \boldsymbol{\theta}) = \sum_{\mathbf{y}_{ij} \in \mathcal{Y}} e^{s(\mathbf{x}_{ij}, \mathbf{y}_{ij}; \boldsymbol{\theta})}$ is the normalizing partition function ($\mathcal{Y}$ is a set of all possible output configurations), and $\boldsymbol{\theta}$ are the parameters of the *score function* (or the negative energy)[1]. By assuming the linear-chain model with *node* cliques ($r \in V$) and *edge* cliques ($e = (r, s) \in E$), the score function $s(\mathbf{x}_{ij}, \mathbf{y}_{ij}; \boldsymbol{\theta})$ can be expressed as the sum:

$$s(\mathbf{x}_{ij}, \mathbf{y}_{ij}; \boldsymbol{\theta}) = \sum_{r \in V} \mathbf{v}^\top \boldsymbol{\Psi}_r^{(V)}(\mathbf{x}_{ij}, y_{ijr}) + \sum_{e=(r,s) \in E} \mathbf{u}^\top \boldsymbol{\Psi}_e^{(E)}(\mathbf{x}_{ij}, y_{ijr}, y_{ijs}), \tag{3}$$

where $\boldsymbol{\theta} = \{\mathbf{v}, \mathbf{u}\}$ are parameters of node features, $\boldsymbol{\Psi}_r^{(V)}(\mathbf{x}_{ij}, y_{ijr})$[2], and edge features, $\boldsymbol{\Psi}_e^{(E)}(\mathbf{x}_{ij}, y_{ijr}, y_{ijs})$, respectively. The score function in (3) has a great modeling flexibility, allowing the node and edge features to be chosen depending on target task.

Node Features. In the CORF model [7], the node features are defined using the homoscedastic ordinal regression model [25] (i.e., with the constant variance σ) as:

$$\begin{aligned} &\mathbf{v}^T \boldsymbol{\Psi}_r^{(V)}(\mathbf{x}_{ij}, y_{ijr}) \\ &\quad \rightarrow \sum_{c=1}^{R} I(y_{ijr} = c) \cdot \left[\Phi\left(\frac{b_{y_{ijr}} - f(x_{ijr})}{\sigma} \right) - \Phi\left(\frac{b_{y_{ijr}-1} - f(x_{ijr})}{\sigma} \right) \right], \end{aligned} \tag{4}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, $I(\cdot)$ is the indicator function that returns 1(0) if the argument is true (false),

[1] For simplicity, we often drop the dependency on $\boldsymbol{\theta}$ in notations.

[2] Unless stated otherwise, we also sometimes drop the time index $r = 1, \ldots T_j$ in x_{ijr} and y_{ijr}

and σ is usually set to 1 for the model identification purpose. In ordinal regression, the difference between the CDFs in (4) is the probability of the observed features, given by x_{ijr}, belonging to class $y_{ijr} = c \in \{1, ..., R\}$ iff $b_{c-1} < f(x_{ijr}) \leq b_c$, where $b_0 = -\infty \leq \cdots \leq b_R = \infty$ are (strictly increasing) thresholds or cut points.

In the standard CORF model [7], $f(x_{ijr}) = \beta x_{ijr}$, where β is the (linear) ordinal projection. In our personalized CORF model, instead of modeling the observed features x_{ijr}, we define this functional form on the identity and AU-intensity spaces, resulting in

$$f(x_{ijr}) \approx \beta_F h_i + \beta_G w_{ijr}, \tag{5}$$

where (h_i, w_{ijr}) are obtained by applying the proposed personalized facial feature algorithm to x_{ijr}, and β_F and β_G are the person- and AU-intensity-specific ordinal projections. By setting $\beta_F = 0$, we obtain the generic CORF model. We use this setting in our experiments for comparisons with the personalized CORF model. Note also that the two-factor functional form in (5) is similar to that in the context-sensitive CORF (cs-CORF) model [15]. However, the key difference is that we define the ordinal projections on different subspaces (the identity and facial AU intensity subspace), while cs-CORF defines the ordinal projections on the single subspace obtained by PCA.

Edge Features. The edge features are defined using the transition model as in the standard CRF:

$$\mathbf{\Psi}_e^{(E)}(y_{ijr}, y_{ijs}) = \Big[I(y_{ijr} = k \ \wedge \ y_{ijs} = l) \Big]_{R \times R}, \tag{6}$$

enforcing the smoothness of the predicted AU intensity levels along the sequence.

Learning and Inference. Using the node and edge features defined above, we arrive at the regularized objective function of the personalized CORF model:

$$\arg\min_{\theta} \sum_{i=1..N_s, j=1..N_i} - \ln P(\mathbf{y}_{ij} | f(\mathbf{x}_{ij}), \theta) + \Omega(\theta), \tag{7}$$

where $\theta = \{b_1, \ldots, b_{R-1}, \mathbf{u}, \beta_F, \beta_G\}$ are the model parameters, and $\Omega(\theta) = \rho_1 \|\mathbf{u}\|^2 + \rho_2 \|\beta_F\|^2 + \rho_3 \|\beta_G\|^2$, is the L_2 regularization used to avoid overfitting of the model parameters. The weights for each term in the regulizer are found using a cross-validation procedure based on a grid search. To ensure that the threshold parameters b satisfy the ordinal constraints, the displacement variables δ_l are introduced, where $b_l = b_1 + \sum_{n=1}^{l-1} \delta_n^2$ for $l = 2, \ldots, R-1$. The quasi-Newton limited-memory BFGS method can then be used to find new (unconstrained) parameters $\theta = \{b_1, \delta_1, \ldots, \delta_{R-2}, \mathbf{u}, \beta_F, \beta_G\}$. Once the model parameters are estimated, inference of test sequences is carried out using Viterbi decoding. For more details about learning and inference in CORF, see [7].

3 Experiments

Evaluation of the proposed approach is performed on the Denver Intensity of Spontaneous Facial Actions (DISFA) dataset [26], the recently published dataset of naturalistic facial AUs that are FACS coded in terms of their intensity using the ordinal scores: 0 (not present) to 5 (maximum intensity). This dataset consists of video recordings of

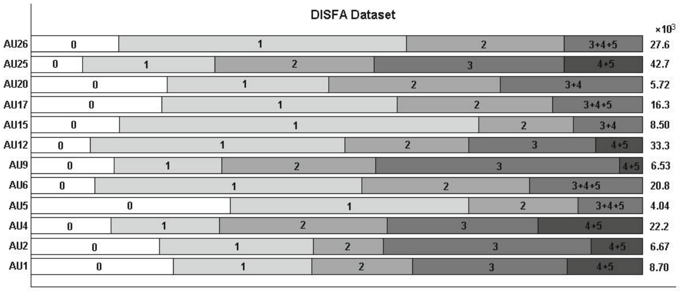


Fig. 1. Distribution of the processed intensity levels for AUs from the DISFA dataset

27 persons while watching short video clips from 'YouTube', resulting in 4845 frames per person. Each image frame was intensity coded in terms of 12 AUs (1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25 and 26). Since most sequences are dominated by the 'neutral' intensity (i.e., 0 level) of AUs, they were pre-segmented per AU. This was attained by pre-segmenting parts of the sequences containing non-neutral intensity by adding the surrounding neutral-intensity frames at the tails of the non-neutral intensity segments. The number of the'neutral' frames was balanced with the second most frequent intensity level of target AUs. Even after this, the coded AU intensity levels were highly imbalanced in the number of their examples, so we grouped the labels of some intensity levels in order to balance the dataset. This is depicted in Fig.1.

As input features we used locations of 66 facial landmarks (see Fig.2) provided by the database creators, and obtained using a 2D Active Appearance Model (2D-AAM) [27]. In the pre-processing step, the facial points were aligned to the corresponding reference face (the average face from the dataset) by applying an affine transform. This is in order to reduce effects of out-of-plane head-rotations. For comparisons of the personalized vs. generic modeling, we formed the following feature sets. First, we processed the aligned facial points by applying PCA to examples of each AU. This resulted, on average, in 20-dimensional feature vectors (preserving 98% of the variance). Following the approach in cs-CORF[15], we subtracted the features of the first frame in each training sequence (with the neutral intensity) from the rest in the sequence, resulting in the feature set denoted as F1=$[x_{ijk}^{pca} - x_{ij1}^{pca}]$ (aka the context-free covariates [15]), and the personalized features, denoted as PF1=$[x_{ijk}^{pca} - x_{ij1}^{pca}; x_{ij1}^{pca}]$, obtained by concatenating the features of the first frame (aka the context covariates [15]) to F1. We then computed the facial features by applying our personalized facial feature decomposition approach to the aligned facial points. For this, the size of the identity space D_h and the AU-intensity-related space D_w varied from 5-15 and 20-25, respectively. These were selected for each AU separately using a validation procedure on the training dataset. The features composed of both the identity and AU-intensity-related spaces are denoted as PF2=$[h_i; w_{ijk}]$, and only of the latter as F2=$[w_{ijk}]$. These features were then used as input to the personalized (using PF1 and PF2, i.e., $\beta_F, \beta_G \neq 0$) and generic (using F1

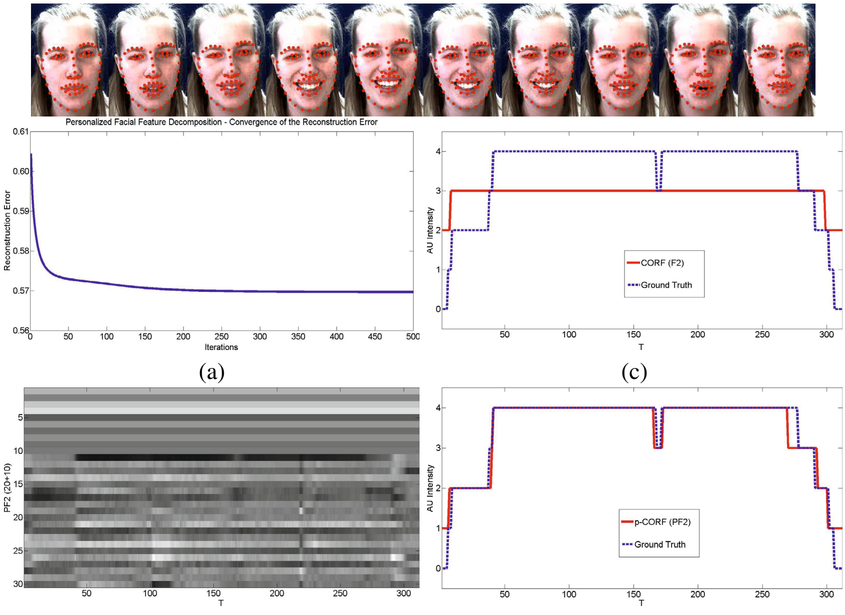
Table 1. The performance of the personalized (PF1/PF2) and generic (F1/F2) CORF models on the task of facial AU intensity estimation using the DISFA dataset. The numbers in *bold* indicate that the personalized CORF (p-CORF) with the proposed PF2 features outperforms the other models on most AUs.

		Upper Face AUs						Lower Face AUs						
		AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	Average
F-1	F1 (CORF[7])	31.7	23.5	29.7	43.8	34.3	31.0	38.4	38.9	39.6	35.8	47.4	31.7	35.5
	PF1 (cs-CORF[15])	**33.8**	24.9	30.2	45.4	40.8	**38.6**	44.2	41.5	40.6	33.0	52.1	37.5	38.5
	F2 (CORF[7])	30.0	23.4	31.2	41.7	37.5	31.4	41.5	42.4	42.0	38.7	53.1	39.2	37.7
	PF2 (p-CORF)	30.4	**26.9**	**33.4**	**46.4**	**44.2**	33.6	**46.7**	**45.0**	**44.3**	**41.6**	**54.6**	**40.4**	**40.6**
MAE	F1 (CORF[7])	1.01	1.08	0.97	0.72	0.80	0.96	0.72	0.77	0.82	0.85	0.63	0.87	0.85
	PF1 (cs-CORF[15])	**0.97**	**1.02**	0.96	0.66	0.73	**0.89**	0.65	0.72	0.80	0.90	0.58	0.81	0.81
	F2 (CORF[7])	1.14	1.11	0.98	0.74	0.76	0.96	0.72	0.73	0.75	0.84	0.54	0.77	0.83
	PF2 (p-CORF)	1.13	1.07	**0.95**	**0.64**	**0.71**	0.92	**0.64**	**0.64**	**0.71**	**0.82**	**0.53**	**0.74**	**0.79**
ICC	F1 (CORF[7])	46.2	44.8	44.8	48.3	36.6	40.1	69.7	44.9	34.1	42.1	68.5	24.7	45.4
	PF1 (cs-CORF[15])	**49.5**	48.1	**45.5**	50.9	44.1	**47.7**	**72.9**	47.8	42.2	36.9	74.0	30.8	49.2
	F2 (CORF[7])	33.8	46.3	41.1	47.6	35.6	42.6	70.4	45.2	46.8	44.4	77.0	36.1	47.2
	PF2 (p-CORF)	41.8	**49.6**	45.1	**53.0**	**45.5**	47.1	71.6	**50.3**	**52.6**	**44.7**	**78.7**	**41.2**	**51.8**

and F2, i.e., $\beta_F = 0, \beta_G \neq 0$) CORF model[3] described in Sec. 2.2. The regularization parameters of the model were selected by a 5-fold cross validation on the training set using a grid-search in the range $\rho = \{10^{-4}, 10^{-3}, ..., 1, 2, 5\}$. If not stated otherwise, in all our experiments we applied a 5-fold cross validation procedure, with each fold containing sequences of different persons. We report accuracy of the models using the average of F-1 scores computed for each intensity level, the mean absolute error (MAE), and Intra-Class Correlation (ICC(3,1)), as reported in [15].

Table 1 shows average results obtained by personalized (PF1/PF2) and generic (F1/F2) CORF models on 12 AUs. We observe that the personalizing CORF model results in an improvement over generic CORF models across all three evaluation scores, evidencing the benefits of using the identity factor for the intensity estimation. We further observe that the personalized CORF (p-CORF) model that uses the proposed facial feature decomposition outperforms the features used in the cs-CORF model. Although this improvement is $\sim 2\%$ for average F-1 score, looking into the results per AU, we see that for some AUs this improvement is considerably larger. For instance, in the case of AU20 (lip stretcher), which involves horizontal motion (elongating the mouth), we achieve an improvement of 8% (F-1). Note that this AU typically occurs in combination with other AUs (e.g., 10+20+25 or 20+26). So, since the identity features in cs-CORF (PF1) are extracted from the first frame in target sequence, the identity component can easily contain the variation due to the co-occurring AUs in that particular sequence. This, in turn, can confuse the model when trying to generalize from such features to other sequences. On the other hand, the personalized features based on the proposed decomposition approach (PF2) include the identity component that is shared across all sequences of target person, thus clearly separating the effects of the identity and AU intensity variation. This is the main reason for better performance attained by the proposed p-CORF. However, there are still some cases when PF1 outperforms the PF2 features, as in the case of AU9 (nose wrinkle). By comparing all three scores, we see that the improvement of PF2 over

[3] In this work, we evaluate the effect of personalizing the base CORF model only [7]. PF1 features correspond to those modeled in cs-CORF[15], however, we do not model the changes in the variance nor we use the weighted max-margin learning as in cs-CORF.



(a)

(c)

(b)

(d)

Fig. 2. The results for AU25(lips part). a) Convergence of the personalized features decomposition. b) Resulting personalized facial features (PF2). The intensity estimation of an example sequence using c) generic CORF (F1), and d) personalized CORF (PF2).

PF1 in F-1 score is more pronounced than in MAE and ICC. This reveals that the model with PF1 missclassifies the neighboring intensity levels more often than when PF2 are used. Again, this indicates that the proposed facial feature decomposition finds the identity component that allows the CORF model to better adapt to each target person.

To further demonstrate the model's performance, in Fig.2 we show learning and inference results for AU25 (lips part). From Fig.2 a), we see that the proposed decomposition algorithm converges in less than 150 iterations. The resulting personalized features (PF2) are depicted in Fig.2 b) using gray-scale. Note that the identity component for the person, the AU intensity of whom we estimate in this example, is constant along the sequence (the upper 10D), while the AU-intensity-specific component varies in time (the lower 20D). From the latter, the change in the features is obvious for the segment covering the time interval from $\sim$ 45-270, in which the AU intensity reaches its peak, as can be seen from the ground truth intensity labels in Fig.2 c),d). By comparing the intensity estimation by generic and personalized CORF, we observe that generic CORF fails to faithfully represent the intensity levels due to over/underestimating of target labels. On the other hand, the personalized CORF almost perfectly fits the target labels, which demonstrates clearly the importance of modeling the identity component.

Finally, Table 2 shows the performance of different generic and personalized models, where the personalization of the models is attained by using the personalized features

Table 2. The average performance of different models on 12 AUs from the DISFA dataset, before and after personalizing the models via the proposed facial features

	F-1				MAE				ICC			
	F1	PF1	F2	PF2	F1	PF1	F2	PF2	F1	PF1	F2	PF2
CORF	35.5	38.5	37.7	**40.6**	0.85	0.81	0.83	**0.79**	45.4	49.2	47.2	**51.8**
CRF	37.3	36.8	37.8	35.7	0.84	0.89	0.87	0.89	41.1	43.1	44.3	37.6
SVM-linear	23.0	23.9	24.4	23.7	0.96	0.99	0.97	1.01	39.2	36.7	36.8	35.3
SVM-RBF	24.7	24.0	23.6	22.1	0.95	0.99	1.00	1.07	41.8	41.4	37.2	33.3

(PF1/PF2). We see from Table 2 that inclusion of the identity component into the traditional classification approaches (CRF and SVM) does not necessarily improve their performance, compared to their generic counterparts (i.e, when F1/F2 are used). Similar observations were also made in [15]. We attribute this to overfitting of the identity component by these models, as its inclusion in both the models, increases the number of parameters to be learned since each class projection is learned independently. This is in contrast to the CORF model, where the ordinal projection is learned jointly for all classes, i.e., intensity levels. Also, the identity component has the central role in adaptation of the model's ordinal thresholds that separate different intensity levels. Thus, for CRF and SVM to take the full benefit of the identity component, these models would need to be formulated so that it is shared among the classes. We plan to investigate this in our future work.

4 Conclusions

In this paper, we proposed an approach for personalized modeling of facial AU intensity. We showed that personalizing the CORF model using the features derived by the proposed personalized facial feature decomposition improves estimation of the AU intensity obtained with generic CORF and other traditional classification models. In our future work, we plan to extend this approach so that it can perform feature decomposition from previously unseen subjects. We also plan to further assess the performance of our approach by comparing it to personalized models based on transfer learning (e.g., [16]) as well as investigate its generalization to high-dimensional facial features (i.e., appearance-based features).

Acknowledgments. This work has been funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA), and the EPSRC grant EP/J017787/1: Analysis of Facial Behaviour for Security in 4D (4D-FAB), and the National Science Foundation under Grant no. IIS0916812.

References

[1] Pantic, M.: Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. Philosophical Transactions of Royal Society B 364, 3505–3513 (2009)

[2] Ekman, P., Friesen, W., Hager, J.: Facial Action Coding System (FACS): Manual. A Human Face (2002)

[3] Mahoor, M., Cadavid, S., Messinger, D., Cohn, J.: A framework for automated measurement of the intensity of non-posed facial action units. In: IEEE CVPR'W, pp. 74–80 (2009)

[4] Vasilescu, M.A.O., Terzopoulos, D.: Multilinear subspace analysis of image ensembles. In: CVPR. vol. 2, p. II-93 (2003)

[5] Mpiperis, I., Malassiotis, S., Strintzis, M.G.: Bilinear models for 3-d face and facial expression recognition. IEEE Trans. on Information Forensics and Security 3, 498–511 (2008)

[6] Valstar, M.F., Pantic, M.: Fully automatic recognition of the temporal phases of facial actions. Systems, Man, and Cybernetics, Part B 42, 28–43 (2012)

[7] Kim, M., Pavlovic, V.: Structured output ordinal regression for dynamic facial emotion intensity prediction. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 649–662. Springer, Heidelberg (2010)

[8] Hablani, R., Chaudhari, N., Tanwani, S.: Recognition of facial expressions using local binary patterns of important facial parts. Int. Journal of Image Proc. 7, 163 (2013)

[9] Arapakis, I., Athanasakos, K., Jose, J.M.: A comparison of general vs personalised affective models for the prediction of topical relevance. In: Proc. of the 33rd Int. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 371–378 (2010)

[10] Dahmane, M., Meunier, J.: Individual feature–appearance for facial action recognition. In: Kamel, M., Campilho, A. (eds.) ICIAR 2011, Part II. LNCS, vol. 6754, pp. 233–242. Springer, Heidelberg (2011)

[11] Doulamis, N.: An adaptable emotionally rich pervasive computing system. In: European Signal Processing Conference (EUSIPCO) (2006)

[12] Romera-Paredes, B., Aung, M.S., Pontil, M., Bianchi-Berthouze, N., de C Williams, A., Watson, P.: Transfer learning to account for idiosyncrasy in face and body expressions. In: FG, pp. 1–6 (2013)

[13] Chen, J., Liu, X., Tu, P., Aragones, A.: Learning person-specific models for facial expression and action unit recognition. Pattern Recognition Letters 34, 1964–1970 (2013)

[14] Romera-Paredes, B., Argyriou, A., Berthouze, N., Pontil, M.: Exploiting unrelated tasks in multi-task learning. In: Int. Conf. on Artificial Intelligence and Statistics, pp. 951–959 (2012)

[15] Rudovic, O., Pavlovic, V., Pantic, M.: Context-sensitive Dynamic Ordinal Regression for Intensity Estimation of Facial Action Units. IEEE TPAMI (in press, 2014)

[16] Chu, W.S., Torre, F.D.L., Cohn, J.F.: Selective transfer machine for personalized facial action unit detection. In: CVPR, pp. 3515–3522 (2013)

[17] Reilly, J., Ghent, J., McDonald, J.: Investigating the dynamics of facial expression. In: Bebis, G., et al. (eds.) ISVC 2006. LNCS, vol. 4292, pp. 334–343. Springer, Heidelberg (2006)

[18] Savrana, A., Sankur, B., Bilgeb, M.: Regression-based intensity estimation of facial action units. In: Image and Vision Computing (2012)

[19] Vasilescu, M.A.O., Terzopoulos, D.: Multilinear image analysis for facial recognition. In: ICPR, vol. 2, p. 20511 (2002)

[20] Wang, H., Ahuja, N.: Facial expression decomposition. In: ICCV, pp. 958–965 (2003)

[21] Lee, C.-S., Elgammal, A.: Facial expression analysis using nonlinear decomposable generative models. In: Zhao, W., Gong, S., Tang, X. (eds.) AMFG 2005. LNCS, vol. 3723,pp. 17–31. Springer, Heidelberg (2005)

[22] Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)

[23] Li, P., Fu, Y., Mohammed, U., Elder, J., Prince, S.: Probabilistic models for inference about identity. TPAMI 34, 144–157 (2012)

[24] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML, pp. 282–289 (2001)

[25] Winkelmann, R., Boes, S.: Analysis of microdata. Springer (2006)

[26] Mavadati, S., Mahoor, M., Bartlett, K., Trinh, P., Cohn, J.: Disfa: A spontaneous facial action intensity database. IEEE Trans. on Affective Comp. 4(2), 151–160 (2013)

[27] Lucey, P., Cohn, J., Prkachin, K., Solomon, P., Matthews, I.: Painful data: The unbc-mcmaster shoulder pain expression archive database. In: IEEE FG, pp. 57–64 (2011)