Gaussian Process Domain Experts for Model Adaptation in Facial Behavior Analysis

Stefanos Eleftheriadis* Ognjen Rudovic* Marc P. Deisenroth* Maja Pantic*†
*Department of Computing, Imperial College London, UK
†EEMCS, University of Twente, The Netherlands
{s.eleftheriadis, orudovic, m.deisenroth, m.pantic}@imperial.ac.uk

Abstract

We present a novel approach for supervised domain adaptation that is based upon the probabilistic framework of Gaussian processes (GPs). Specifically, we introduce domain-specific GPs as local experts for facial expression classification from face images. The adaptation of the classifier is facilitated in probabilistic fashion by conditioning the target expert on multiple source experts. Furthermore, in contrast to existing adaptation approaches, we also learn a target expert from available target data solely. Then, a single and confident classifier is obtained by combining the predictions from multiple experts based on their confidence. Learning of the model is efficient and requires no retraining/reweighting of the source classifiers. We evaluate the proposed approach on two publicly available datasets for multi-class (MultiPIE) and multi-label (DISFA) facial expression classification. To this end, we perform adaptation of two contextual factors: 'where' (view) and 'who' (subject). We show in our experiments that the proposed approach consistently outperforms both source and target classifiers, while using as few as 30 target examples. It also outperforms the state-of-the-art approaches for supervised domain adaptation.

1. Introduction

Human face is believed to be the most powerful channel for conveying, non-verbally, behavioral traits such as personality, intentions and affect, among others [2, 34]. Facial expressions can be studied at the message level (interpretation in terms of the message conveyed, *e.g.*, emotions), and sign level (analysis of facial muscle movements named action units (AUs)). To this end, the Facial Action Coding System (FACS) [10] has been used. It is the most comprehensive anatomically-based system for describing facial expressions at both the levels. FACS defines 33 unique AUs, and several categories of head/eye movements.

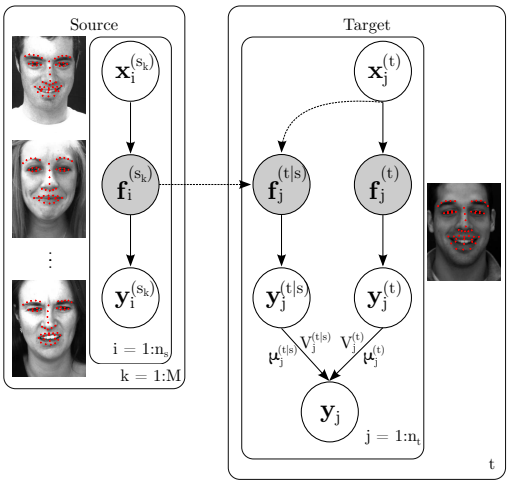


Figure 1. The proposed GPDE model. The learning consists of training the multiple source ($s_k, k = 1, \cdots, M$) and the target (t) GP experts (in this case, each subject is treated as an expert), using the available labeled training data pairs ($\boldsymbol{x}, \boldsymbol{y}$) – the input features (*e.g.*, facial landmarks) and output labels (*e.g.*, AU activations), respectively. Adaptation (dashed lines) for the target data is performed via conditioning the latent functions, $\boldsymbol{f}$, of the target GP on the source experts ($t|s$). During inference, we fuse the predictions from the experts ($\boldsymbol{\mu}^{\{t,(t|s)\}}$) by means of their predictive variance ($V^{\{t,(t|s)\}}$), with the role of a confidence measure.

Due to its practical importance in medicine, marketing and entertainment, automated analysis of facial expressions has received significant research attention over the last two decades. Despite rapid advances in computer vision and machine learning, majority of the models proposed so far for facial expression analysis rely on generic classifiers. These classifiers are expected to generalize well when applied to data recorded within specific contexts, as defined by the W5+ context questions ('who', 'where', 'how', 'what', 'when' and 'why') [27]. Nevertheless, due to possible variations in these contextual dimensions, the performance of virtually all existing generic classifiers for facial expression analysis is expected to downgrade largely when applied to previously unseen data [12]. This

is especially pronounced in the case of unseen subjects (due to variation in their age, gender, expressiveness), changes in pose and illumination, environments, and so on. To circumvent these challenges, two lines of work have been proposed. The first relies on carefull design of 'context-independent' image features and the use of generic classifiers [35, 17, 23], while the second attempts adaptation of the target classifiers [27, 5, 29]. In this work, we employ the latter approach and focus on adaptation of the context questions 'where' and 'who' in our data.

Variation in head-pose and illumination ('where') has been addressed by combining illumination invariant features with multi-view learning techniques [35, 17, 23, 26, 14, 11]. On the other hand, the individual differences among subjects ('who') have mainly been tackled by accounting for the subject information at the training stage. Specifically, the original feature set is extended by adding the subject-specific features [27], or by building person-specific classifiers [31]. Although these approaches showed improvement over generic classifiers, there is still a number of challenges to address. In particular, the multi-view learning requires a large amount of images in various poses, which is typically not available. On the other hand, for building personalized classifiers, access to an adequate collection of images of the target person is essential. Consequently, existing approaches perform re-weighting previosly learned classifiers to fit the target data (*e.g.*, [5]), or training of new models using the additional target data. However, both of these are sub-optimal. Thus, our aim is to find an effective approach to adapt the already trained generic models for facial behavior analysis by using a small number of target data. In the case of the context question 'where', this boils down to adapting the frontal classifier to a non-frontal view using only a small number of expressive images from the target view. Similarly, in the case of the subject adaptation ('who'), the model adaptation is performed by using as few annotated images of target subject as needed to gain in the prediction performance (*e.g.*, AU detection). This approach is expected to generalize better than generic classifiers learned from the available source and/or target (training) data.

To address the challenges mentioned above, we use the notion of *domain adaptation* to perform two tasks: (i) view and (ii) subject adaptation, for facial expression recognition (FER) and AU detection. In particular, we address the problem of domain adaptation where the distribution of the (facial) features varies across domains (*i.e.*, contexts such as the view or subject), while the output labels (in our case, the emotion or AU activations) remain the same. This is also known as *covariate shift*, and the two domains are called *source* (*e.g.*, frontal view) and *target* (*e.g.*, profile view) domain, respectively. Furthermore, a *supervised* setting, where a small number of labeled target examples is available during the adaptation process, is assumed. We build our model upon the probabilistic framework of Gaussian processes (GPs) [25], and generalize the product of expert models [7, 3] to domain adaptation scenario. More specifically, instead of adjusting the classifier parameters between the domains, as in [5, 33, 4, 22, 29], we propose domain-specific GP experts that model the domain specific data.[1] Moreover, instead of minimizing the error between the distributions of the original source and target domain data [5, 22], we use Bayesian domain adaptation [20] and explain the target data by conditioning on the learned source experts. An advantage of our probabilistic formulation is that during adaptation, we exploit the variance in the predictions when combining the source and target domains [30]. This results in a *confident* classifier that minimizes the risk of potential negative transfer (*i.e.*, the adapted model performing worse than the model trained using the adaptation data only). In contrast to transductive adaptation approaches (*e.g.*, [5]) that need to be re-trained completely, adaptation of our model is efficient and requires no re-training of the source model. The model outline is depicted in Fig. 1. The contributions of this work can be summarized as follows:

- We present a novel approach for supervised domain adaptation that can, for the first time, perform adaptation to contextual factors 'where' (across different views) and 'who' (by personalizing the target classifier) during modeling of facial expression data.
- To the best of our knowledge, this is the first work in the domain of facial behavior modeling that can simultaneously perform adaption to multiple outputs (*i.e.*, AUs). Existing models in the field that attempt the model adaptation do so for each output independently.
- Due to its probabilistic nature, the proposed approach provides the confidence in the predicted labels for the target expressions. This is in contrast to majority of the models that are purely discriminative, and thus, cannot provide a measure of how 'reliable' the predictions are.
- We show in our experiments on view and subject adaptation that the proposed model can generalize better than source and target domains together by using as few as 30 target samples to perform the adaptation. Furthermore, virtually all existing domain adaptation approaches fail to reach the performance of the target classifiers when more target data become available (negative transfer). Our approach overcomes this due to the newly introduced scheme for combining the source and target experts.

[1] The use of GPs for this task is motivated by their good generalization abilities, even when trained with limited amount of data [25, 11]. This property is crucial for the training of the target expert, since the available data are scarce.

2. Related Work

2.1. Domain Adaptation in Facial Behavior Analysis

The majority of approaches for domain adaptation in the context of facial behavior analysis focus on building personalized classifiers for the test subjects. For instance, [22] uses the supervised kernel mean matching (KMM) to align the source and target data distributions. This is achieved by re-weighting the source data, which, in combination with the target data, form the input features that are used to train the support vector machine (SVM) [6] classifier for FER. Likewise, [5] uses unsupervised KMM to learn person-specific AU detectors. This is attained by modifying the SVM cost function to account for the KMM between source and target data, adjusting the SVM's hyperplane to the target test data. However, this results in the transductive learning approach, thus, the classifier has to be re-learned for each target subject. In [4], a two-step learning approach is proposed for person-specific pain recognition and AU detection. First, data of each subject are regarded as different source domains, and are used to train weak Adaboost classifiers. Then, the weak classifiers are weighted based on their classification performance on the available target data. In [29, 5], the Adaboost classifiers are replaced with the linear SVMs, and then the support vector regression (SVR) is employed to learn the mapping from the feature distribution to the parameters of the SVM classifier.

Note that, apart from [4], all the works mentioned above perform in the unsupervised adaptation setting. While this requires less effort in terms of obtaining the labels for the target sub-sample, its underlying assumption is that target data can be well represented as a weighted combination of the source data. However, in real-world data, this assumption can easily be violated, resulting in poor performance of the adapted classifier. In this work, we adopt a supervised approach that needs only a few annotated data from target domain to perform the adaptation. This, in turn, allows us to define both target and source experts, assuring that the performance of the resulting classifier is not constrained by the distribution of the source data, as in unsupervised adaptation approaches. Contrary to transductive learning approaches such as [5], our approach requires adaptation of the target expert solely, without the need to re-learn the source experts, resulting in an efficient adaptation process. Moreover, in contrast to our approach, none of the aforementioned works provides a measure of confidence in the predicted labels. Finally, note that the proposed approach and the methods mentioned above differ from those recently proposed for transfer learning [1]. The goal of the latter is to adapt a classifier learned for, *e.g.*, one AU to another AU, which is different from the adaptation task addressed in this work.

2.2. Domain Adaptation

Domain adaptation is a well studied problem in machine learning (for an extensive survey, see [24]). Here we review relevant (semi-)supervised adaptation approaches. For instance, [19] learns a transformation that maximizes similarity between data in the source and target domains by enforcing data pairs with the same labels to have high similarity, and pairs with different labels to be dissimilar. Then, a k-NN classifier is used to perform classification of target data. [15] is an extension of this approach to multiple source domains. The input data are assumed to be generated from category-specific local domain mixtures, the mixing weights of which determine the underlying domain of the data, classified using an SVM classifier. Similarly, [16] learns a linear asymmetric transformation to maximally align target features to the source domain. This is attained by introducing max-margin constraints that allow the learning of the transformation matrix and SVM classifier jointly. [8] extends the work in [16] by introducing additional constraints to the max-margin formulation. More specifically, unlabeled data from the target domain are used to enforce the classifier to produce similar predictions for similar target-source data. While these methods attempt to directly align the target to source features, several works attempted this through a shared manifold. For instance, [9] learns a non-linear transformation from both source and target data to a shared latent space, along with the target classifier. Likewise, [32] finds a low-dimensional subspace, which preserves the structure across the domains. The subspace is facilitated by projection functions that are learned jointly with the linear classifier. Again, the structure preservation constraints are used to ensure that similar data across domains are close in the subspace.

All of the above methods tackle the adaptation problem in a deterministic fashion, thus they do not provide a measure of confidence in the target predictions. By contrast, our approach is fully probabilistic and non-parametric due to the use of GPs. The proposed is related to recent advances in the GP literature [20, 18] on domain adaptation. Specifically, in [20], the predictive distribution of a GP trained on the source data is used as a prior for making inference in the target domain. Similarly, [18] proposed a two-layer GP that jointly learns separate discriminative functions from the source and target features to the labels. The intermediate layer facilitates the adaptation step, and variational approximation is employed to integrate out this layer. In contrast to [20], the proposed defines a target specific expert, which is then combined in a principled manner with the source domain experts. The benefit of this is that the resulting classifier is not limited by the distribution of the source data. Also, in contrast to [18], the training of the experts is performed independently, and thus, we need not retrain the source classifier.

3. Problem Formulation

We consider a supervised setting for domain adaptation, where we have access to a large collection of labeled *source* domain data, and a smaller set of labeled *target* domain data. Let $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ be the input (features) and output (labels) spaces, respectively. We assume that the input space is comprised of the source and target domains, $\mathcal{S}$ and $\mathcal{T}$, respectively, that may differ in feature distribution. Hence, $\boldsymbol{X}^{(s)} = \{\boldsymbol{x}_{n_s}^{(s)}\}_{n_s=1}^{N_s}$ and $\boldsymbol{X}^{(t)} = \{\boldsymbol{x}_{n_t}^{(t)}\}_{n_t=1}^{N_t}$, with $\boldsymbol{x}_{n_s}^{(s)}, \boldsymbol{x}_{n_t}^{(t)} \in \mathbb{R}^D$, and $N_t \ll N_s$. In our case, these can be different views or subjects. On the other hand, $\boldsymbol{Y}^{(s)} = \{\boldsymbol{y}_{n_s}^{(s)}\}_{n_s=1}^{N_s}$ and $\boldsymbol{Y}^{(t)} = \{\boldsymbol{y}_{n_t}^{(t)}\}_{n_t=1}^{N_t}$ correspond to same labels for both source and target domains. Each vector $\boldsymbol{y}_n^{\{s,t\}}$ contains the binary class labels of C classes. We now formulate the regression problem as:

$$\boldsymbol{y}_{n_v}^{(v)} = f^{(v)}(\boldsymbol{x}_{n_v}^{(v)}) + \epsilon^{(v)}, \tag{1}$$

where $\epsilon^{(v)} \sim \mathcal{N}(0, \sigma_v^2)$ is i.i.d. additive Gaussian noise, and the index $v \in \{s, t\}$ denotes the dependence on each domain. The objective is to infer the latent functions $f^{(v)}$, given the training dataset $\mathcal{D}^{(v)} = \{\boldsymbol{X}^{(v)}, \boldsymbol{Y}^{(v)}\}$. By following the framework of GPs [25], we place a prior on the functions $f^{(v)}$, so that the function values $\boldsymbol{f}_{n_v}^{(v)} = f^{(v)}(\boldsymbol{x}_{n_v}^{(v)})$ follow a Gaussian distribution $p(\boldsymbol{F}^{(v)}|\boldsymbol{X}^{(v)}) = \mathcal{N}(\boldsymbol{F}^{(v)}|\boldsymbol{0}, \boldsymbol{K}^{(v)})$. Here, $\boldsymbol{F}^{(v)} = \{\boldsymbol{f}_{n_v}^{(v)}\}_{n_v=1}^{N_v}$, and $\boldsymbol{K}^{(v)} = k^{(v)}(\boldsymbol{X}^{(v)}, \boldsymbol{X}^{(v)})$ is the kernel covariance function, which is assumed to be shared among the label dimensions. In this work, we use the radial basis function (RBF) kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\Big(-\frac{1}{2\ell^2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\Big), \tag{2}$$

where $\{\ell, \sigma_f\}$ are the kernel hyper-parameters. The regression mapping can be fully defined by the set of hyper-parameters $\boldsymbol{\theta} = \{\ell, \sigma_f, \sigma_v\}$. Training of the GP consists of finding the hyper-parameters that maximize the log-marginal likelihood

$$\log p(\boldsymbol{Y}^{(v)}|\boldsymbol{X}^{(v)}, \boldsymbol{\theta}^{(v)}) = -\frac{1}{2}\text{tr}\Big[(\boldsymbol{K}^{(v)} + \sigma_v^2\boldsymbol{I})^{-1}\boldsymbol{Y}^{(v)}\boldsymbol{Y}^{(v)^T}\Big] - \frac{C}{2}\log|\boldsymbol{K}^{(v)} + \sigma_v^2\boldsymbol{I}| + \text{const.} \tag{3}$$

Given a test input $\boldsymbol{x}_*^{(v)}$ we obtain the GP predictive distribution by conditioning on the training data $\mathcal{D}^{(v)}$ as $p(\boldsymbol{f}_*^{(v)}|\boldsymbol{x}_*^{(v)}, \mathcal{D}^{(v)}) = \mathcal{N}(\mu^{(v)}(\boldsymbol{x}_*^{(v)}), V^{(v)}(\boldsymbol{x}_*^{(v)}))$ with

$$\mu^{(v)}(\boldsymbol{x}_*^{(v)}) = \boldsymbol{k}_*^{(v)^T}(\boldsymbol{K}^{(v)} + \sigma_v^2\boldsymbol{I})^{-1}\boldsymbol{Y}^{(v)} \tag{4}$$

$$V^{(v)}(\boldsymbol{x}_*^{(v)}) = k_{**}^{(v)} - \boldsymbol{k}_*^{(v)^T}(\boldsymbol{K}^{(v)} + \sigma_v^2\boldsymbol{I})^{-1}\boldsymbol{k}_*^{(v)}, \tag{5}$$

where $\boldsymbol{k}_*^{(v)} = k^{(v)}(\boldsymbol{X}^{(v)}, \boldsymbol{x}_*^{(v)})$ and $k_{**}^{(v)} = k^{(v)}(\boldsymbol{x}_*^{(v)}, \boldsymbol{x}_*^{(v)})$. For convenience we denote $\boldsymbol{\mu}_*^{(v)} = \mu^{(v)}(\boldsymbol{x}_*^{(v)})$ and $V_{**}^{(v)} = V^{(v)}(\boldsymbol{x}_*^{(v)})$. Within the introduced notation, we have the choice to learn either (i) independent functions $f^{(v)}$ or (ii) a universal function f that couples the data from the two domains. However, neither option allows us to explore the idea of domain adaptation: In the former we learn domain-specific models, while in the latter we simplify the problem by concatenating the data from the two domains.

4. Domain Conditioned GPs

4.1. GP Adaptation

A straightforward approach to obtain a model capable of performing inference on data from both domains is to assume the existence of a universal latent function with a single set of hyper-parameters $\boldsymbol{\theta}$. To this end, the authors in [20] proposed a simple, yet effective, three-step approach for GP adaptation (GPA):

1. Train a GP on the source data with likelihood $p(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)}, \boldsymbol{\theta})$ to learn the hyper-parameters $\boldsymbol{\theta}$. The posterior distribution is the given by Eqs. (4–5).
2. Use the obtained posterior distribution of the source data, as a prior for the GP of the target data $p(\boldsymbol{Y}^{(t)}|\boldsymbol{X}^{(t)}, \mathcal{D}^{(s)}, \boldsymbol{\theta})$.
3. Correct the posterior distribution to account for the target data $\mathcal{D}^{(t)}$ as well.

The prior of the target data in the second step is given by applying Eqs. (4–5) on $\boldsymbol{X}^{(t)}$

$$\boldsymbol{\mu}^{(t|s)} = \boldsymbol{K}_{st}^{(s)^T}(\boldsymbol{K}^{(s)} + \sigma_s^2\boldsymbol{I})^{-1}\boldsymbol{Y}^{(s)} \tag{6}$$

$$\boldsymbol{V}^{(t|s)} = \boldsymbol{K}_{tt}^{(s)} - \boldsymbol{K}_{st}^{(s)^T}(\boldsymbol{K}^{(s)} + \sigma_s^2\boldsymbol{I})^{-1}\boldsymbol{K}_{st}^{(s)}, \tag{7}$$

where $\boldsymbol{K}_{tt}^{(s)} = k^{(s)}(\boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t)}), \boldsymbol{K}_{st}^{(s)} = k^{(s)}(\boldsymbol{X}^{(s)}, \boldsymbol{X}^{(t)})$, and the superscript $t|s$ denotes the conditioning order. Given the above prior and a test input $\boldsymbol{x}_*^{(t)}$, the correct form of the adapted posterior after observing the target domain data is given by:

$$\mu_{ad}^{(s)}(\boldsymbol{x}_*^{(t)}) = \boldsymbol{\mu}_*^{(s)} + \boldsymbol{V}_*^{(t|s)^T}(\boldsymbol{V}^{(t|s)} + \sigma_s^2\boldsymbol{I})^{-1}(\boldsymbol{Y}^{(t)} - \boldsymbol{\mu}^{(t|s)}) \tag{8}$$

$$V_{ad}^{(s)}(\boldsymbol{x}_*^{(t)}) = V_{**}^{(s)} - \boldsymbol{V}_*^{(t|s)^T}(\boldsymbol{V}^{(t|s)} + \sigma_s^2\boldsymbol{I})^{-1}\boldsymbol{V}_*^{(t|s)}, \tag{9}$$

with $\boldsymbol{V}_*^{(t|s)} = k^{(s)}(\boldsymbol{X}^{(t)}, \boldsymbol{x}_*^{(t)}) - k^{(s)}(\boldsymbol{X}^{(s)}, \boldsymbol{X}^{(t)})^T(\boldsymbol{K}^{(s)} + \sigma_s^2\boldsymbol{I})^{-1}k^{(s)}(\boldsymbol{X}^{(s)}, \boldsymbol{x}_*^{(t)})$.

Eqs. (8–9) shows that final prediction in the GPA is the combination of the original prediction based on the source data only, plus a correction term. The latter shifts the mean toward the distribution of the target data and improves the model's confidence by reducing the predictive variance. Note that we originally constrained the model to learn a single latent function f for both conditional

distributions $p(\boldsymbol{Y}^{(v)}|\boldsymbol{X}^{(v)})$ to derive the posterior for the GPA. However, this constraint implicitly assumes that the marginal distributions of the data $p(\boldsymbol{X}^{(v)})$ are similar. This assumption violates the general idea of domain adaptation, where by definition, the marginals may have significantly different attributes (*e.g.*, input features from different observation views). In such cases, GPA could perform worse than an independent GP trained solely on the target data $\mathcal{D}^{(t)}$. One possible way to address this issue is to retrain the $\log p(\boldsymbol{Y}^{(t)}|\boldsymbol{X}^{(t)}, \mathcal{D}^{(s)}, \boldsymbol{\theta})$ of the GPA w.r.t. $\boldsymbol{\theta}$ [20]. This option will compensate for the differences in the distributions by readjusting the hyper-parameters. However, it comes with the price of retraining of the model. Furthermore, it does not allow for modeling domain-specific attributes since the predictions are still determined mainly from the source distribution.

4.2. GP Domain Experts

Product of GP Experts. In the proposed approach, we assume that each expert is a GP that operates only on a subset of data, *i.e.*, $\mathcal{D}^{(s)}, \mathcal{D}^{(t)}$. Hence, we can follow the methodology presented in Sec. 3 in order to train domain-specific GPs and learn different latent functions, *i.e.*, hyper-parameters $\boldsymbol{\theta}^{(v)}$. Within the current formulation we treat the source domain as a combination of multiple source datasets (*e.g.*, subject-specific datasets) $\mathcal{D}^{(s)} = \{\mathcal{D}^{(s_1)}, \ldots, \mathcal{D}^{(s_M)}\}$, where M is the total number of source domains (datasets).

Training. Given the above mentioned data split and assuming conditional independence, the marginal likelihood can be approximated by

$$p(\boldsymbol{Y}^{\{s,t\}}, |\boldsymbol{X}^{\{s,t\}}, \boldsymbol{\theta}^{\{s,t\}}) = p(\boldsymbol{Y}^{(t)}|\boldsymbol{X}^{(t)}, \boldsymbol{\theta}^{(t)}) \prod_{k=1}^{M} p_k(\boldsymbol{Y}^{(s_k)}|\boldsymbol{X}^{(s_k)}, \boldsymbol{\theta}^{(s)}). \quad (10)$$

Note that we share the set of hyper-parameters $\boldsymbol{\theta}^{(s)}$ across all the source domains. The intuition behind this is that in each source domain we may observe different label distribution $p(\boldsymbol{Y}^{(s_k)})$, yet after exploiting all the available datasets we can model the overall distribution $p(\boldsymbol{Y}^{(s)})$ with a single set of hyper-parameters $\boldsymbol{\theta}^{(s)}$. However, this does not guarantee that we are also able to explain the target label distribution $p(\boldsymbol{Y}^{(t)})$ with the same hyper-parameters. Thus, we also search for $\boldsymbol{\theta}^{(t)}$ for modeling the domain-specific attributes. Similarly to Sec. 3 learning of the hyper-parameters is performed by maximizing

$$\log p(\boldsymbol{Y}^{\{s,t\}}, |\boldsymbol{X}^{\{s,t\}}, \boldsymbol{\theta}^{\{s,t\}}) = \log p(\boldsymbol{Y}^{(t)}|\boldsymbol{X}^{(t)}, \boldsymbol{\theta}^{(t)}) + \sum_{k=1}^{M} \log p_k(\boldsymbol{Y}^{(s_k)}|\boldsymbol{X}^{(s_k)}, \boldsymbol{\theta}^{(s)}), \quad (11)$$

Algorithm 1 Domain adaptation with GPDE

Inputs: $\mathcal{D}^{(s)} = \{\boldsymbol{X}^{(s)}, \boldsymbol{Y}^{(s)}\}, \mathcal{D}^{(t)} = \{\boldsymbol{X}^{(t)}, \boldsymbol{Y}^{(t)}\}$
Training:
Learn the hyper-parameters $\boldsymbol{\theta}^{\{s,t\}}$ by maximizing Eq. (11).
Adaptation:
Adapt the posterior from the source experts via Eq. (8–9).
Predictions of Experts:
Combine the prediction from each GP domain expert via Eq. (13–14).
Output: $\boldsymbol{y}_* = sign(\boldsymbol{\mu}_*^{gpde})$.

where each log-marginal is computed according to Eq. (3). The above factorization, apart from facilitating learning of the domain experts, allows for efficient GP training even with larger datasets, as shown in [7]. Note that the source experts can be learned independently from the target, which allows our model to generalize to unseen target domains without retraining.

Predictions. Once we have trained the GPDE, we need to combine the predictions from each expert to form an overall prediction. To this end, we follow the approach presented in [3], where we further readjust the predictions from the source experts using the trick of GPA. Hence, the predictive distribution is given by

$$p(\boldsymbol{f}_*^{(t)}|\boldsymbol{x}_*^{(t)}, \mathcal{D}) = \prod_{k=1}^{M} p_k^{\beta_{s_k}}(\boldsymbol{f}_*^{(t)}|\boldsymbol{x}_*^{(t)}, \mathcal{D}^{(s_k)}, \mathcal{D}^{(t)}, \boldsymbol{\theta}^{(s)}) \cdot p^{\beta_t}(\boldsymbol{f}_*^{(t)}|\boldsymbol{x}_*^{(t)}, \mathcal{D}^{(t)}, \boldsymbol{\theta}^{(t)}), \quad (12)$$

where β_{s_k}, β_t control the contribution of each expert. In this work we equally weight the experts and normalize them such that $\beta_t + \sum \beta_{s_k} = 1$, as suggested in [7]. The predictive mean and variance are given by

$$\boldsymbol{\mu}_*^{\text{gpde}} = V_*^{\text{gpde}} \left[\beta_t {V_*^{(t)}}^{-1} \boldsymbol{\mu}_*^{(t)} + \sum\nolimits_k \beta_{s_k} {V_{ad}^{(s_k)}}^{-1} \boldsymbol{\mu}_{ad}^{(s_k)}\right] \quad (13)$$

$$V_*^{\text{gpde}} = \left[\beta_t {V_*^{(t)}}^{-1} + \sum\nolimits_k \beta_{s_k} {V_{ad}^{(s_k)}}^{-1}\right]^{-1}. \quad (14)$$

At this point the contribution of the GPDE becomes clear: Eq. (13) shows that the overall mean is the sum of the predictions from each expert, weighted by their precision (inverse variance). Hence, the solution of the GPDE will favor the predictions of more confident experts. On the other hand, if the quality of a domain expert is poor (noisy predictions with large variance), GPDE will weaken its contribution to the overall prediction. Algorithm 1 summarizes the GPDE adaptation procedure.

5. Experiments

We evaluate the proposed model on acted and spontaneous facial expressions from two publicly available

datasets: MultiPIE [13] and Denver Intensity of Spontaneous Facial Actions (DISFA) [21]. Specifically, MultiPIE contains images of 373 subjects depicting acted facial expressions of Neutral (NE), Disgust (DI), Surprise (SU), Smile (SM), Scream (SC) and Squint (SQ), captured at various pan angles. In our experiments, we used images from 0°, -15° and -30°. DISFA is widely used in the AU-related literature, due to the large amount of (subjects and AUs) annotated images. It contains video recordings of 27 subjects while watching YouTube videos. Each frame is coded in terms of the AU intensity on a six-point ordinal scale. For our experiments we selected the six most frequently occurring AUs, *i.e.*, AUs (4, 6, 9, 12, 25, 26), while we treated each AU with intensity larger than zero as active.

Features: We use a set of geometric features derived from the facial landmark locations. DISFA dataset comes with frame-by-frame annotations of 66 facial points, while the same annotated points for MultiPIE were obtained from [28]. We discarded the contour landmarks, leading to the set of 49 facial points. These were then registered to a reference face (average face per view for MultiPIE, and average face for DISFA) using an affine transform. In order to further remove potential noise and artifacts, the aligned landmark points were post-processed via PCA, retaining 99% of the energy, which resulted in 30D feature vectors.

Evaluation procedure. We evaluate GPDE on both multi-class (FER on MultiPIE) and multi-label (multiple AU detection on DISFA) scenarios. We also assess the adaptation capacity of the model with a single (view adaptation) and multiple (subject adaptation) source domains. For the task of FER, the frontal view, *i.e.*, 0°, served as a single source domain, and inference was performed via adaptation to the target domains -15° and -30°. For the AU detection task, the various subjects from the train data were used as multiple source domains, and adaptation was performed each time on the tested subject. To evaluate the model's adaptation ability we strictly follow a training protocol, where for each experiment we vary the cardinality of the training target data (we always use all the available source domain data). For MultiPIE, we first split the data in 5-folds (4 training, 1 testing) and then, we keep increasing the cardinality as: $N_t = 10, 30, 50, 100, 200, 300, 600, 1200$. For DISFA we partition the data in 3-folds (20 training source subjects at a time). From the test subject's sequence the first 500 frames were used as target training data (with increasing cardinality $N_t = 10, 30, 50, 100, 200, 500$), while inference was performed on the rest frames of the sequence. This is in order to avoid the target model overfitting the temporally neighboring examples of test subject. For the FER experiments, we employ the classification ratio (CR) as the evaluation measure, while for the AU detection, due to the imbalance in the data, we report the F1 score and the area under the ROC curve (AUC).

Models compared. We compare the proposed GPDE with the two generic models GP_{source} and GP_{target}. The former is trained solely on the source data, while the latter on the target data used for the adaptation. Furthermore, we compare to the state-of-the-art models for supervised domain adaptation, *i.e.*, the GPA [20] and the asymmetric transfer learning with deep GP (ATL-DGP) [18]. The GPA is an instance of the proposed GPDE, with only a source domain expert (no target) and predictions given by Eq. (8–9). ATL-DGP[2] employs an intermediate GP to combine the predictions of GP_{source} and GP_{target}. In the multi-source experiment we also compare to GPDE_{ss}, which is the instance of GPDE with all the subjects treated as a single source domain. Note that We do not include comparisons with the deterministic approaches (*e.g.*, [16, 19]), as it has been shown in [18] that ATL-DGP outperforms these methods.

5.1. View adaptation from a single source

In this experiment, we demonstrate the effectiveness of the proposed approach when the distributions between source and target domain ($0^\circ \rightarrow -15^\circ$ and $0^\circ \rightarrow -30^\circ$) differ in an increasing non-linear manner. For this purpose we evaluate all considered algorithms in terms of their ability to perform accurate FER as we move away from the frontal pose. Example images for the specified task can be seen in Fig. 2. Table 1 summarizes the results. The generic classifier GP_{source} exhibits the lowest performance, due to the fact that it has only been trained on source domain images. It is important to note the drop in the classification rate ($\approx 5\%$) when the target domain changes from -15° to -30°. This indicates the inefficiency of a generic classifier to deal with data of different characteristics. On the other hand, the GP_{target} when trained with as few as 30 data points achieves similar performance to the GP_{source} since it benefits from modeling domain-specific attributes. A further increase of the cardinality of the target training data results in a significant improve in the classification rate. A similar trend can be observed in the performance of the adaptation methods, where the inclusion of 10 labeled data points from the target domain is adequate to shift the learned source classifier towards the distribution of the target data.

The GPA uses the extra data to condition on the generic classifier GP_{source} and increase its prediction performance. ATL-DGP on the other hand facilitates a joint learning scheme where GP_{source} and GP_{target} are fused together, via conditioning, in a deep architecture. The advantage of the latter is evidenced from the highest achieved accuracy, *i.e.*, 83.32% for $N_t = 10$. However, the joint training scheme of ATL-DGP limits its adaptation ability, due to the high effect of the source prior. Consequently, its per-

[2]The provided code for ATL-DGP is not capable of multi-label classification, since it treats the labels only in a 1-of-K encoding. Thus, it cannot be evaluated on the multiple AU detection task.

Table 1. Average classification rate across 5-folds on MultiPIE. The adaptation is performed from $0^\circ \rightarrow -15^\circ$ and $0^\circ \rightarrow -30^\circ$, with increasing cardinality of labeled target domain data ($10 - 1200$).

Method	$0^\circ \rightarrow -15^\circ$								$0^\circ \rightarrow -30^\circ$							
	10	30	50	100	200	300	600	1200	10	30	50	100	200	300	600	1200
GP_{source}	81.65								76.94							
GP_{target}	55.85	81.19	84.59	89.61	90.66	91.31	91.57	97.26	51.99	76.09	81.97	86.48	88.57	89.75	**92.16**	**98.43**
GPA [20]	82.36	84.00	85.37	88.63	90.20	91.51	93.79	96.15	77.73	79.82	81.65	85.43	87.79	87.72	89.29	93.01
ATL-DGP [18]	**83.32**	86.34	85.22	85.62	85.16	86.42	86.53	87.80	**79.82**	**82.93**	83.36	85.53	82.08	84.32	80.03	83.04
GPDE	82.95	**86.35**	**87.52**	**92.10**	**93.73**	**94.64**	**95.36**	**97.84**	78.71	82.17	**84.65**	**87.85**	**88.83**	**90.01**	91.38	96.86

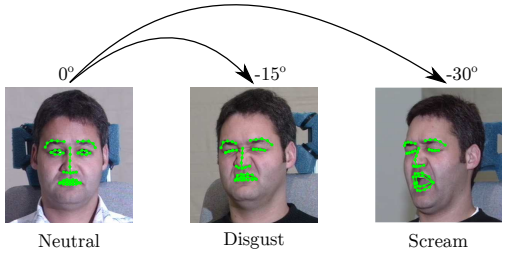


Figure 2. View adaptation for FER on MultiPIE dataset.

formance saturates. A further disadvantage of ATL-DGP's joint learning is that it requires retraining every time the target distribution changes. Finally, the proposed GPDE, uses the notion of experts to unify GP_{source} and GP_{target} into a single classifier. To achieve so, GPDE measures the confidence of the predictions from each expert (by means of predictive variance), in contrast to GPA (uses source expert only) and ATL-DGP (uses an uninformative prior). This property of GPDE is more pronounced in the adaptation $0^\circ \rightarrow -30^\circ$ with $N_t > 300$, where GP_{target} achieves the highest classification ratio. GPDE performs similarly to the target expert while, GPA and ATL-DGP underestimate the prediction capacity of the target-specific classifier, and thus, attain lower results.

A better insight into the performance of the considered methods can be obtained from the confusion matrices in Fig. 3. The reported results are for $0^\circ \rightarrow -30^\circ$ adaptation with $N_t = 50$ (at which point the GP_{target} starts outperforming GP_{source}). The proposed GPDE takes advantage of the target-specific expert and significantly reduces the confusion between the subtle expressions of Disgust and Squint with the Neutral face.

5.2. Subject adaptation from multiple sources

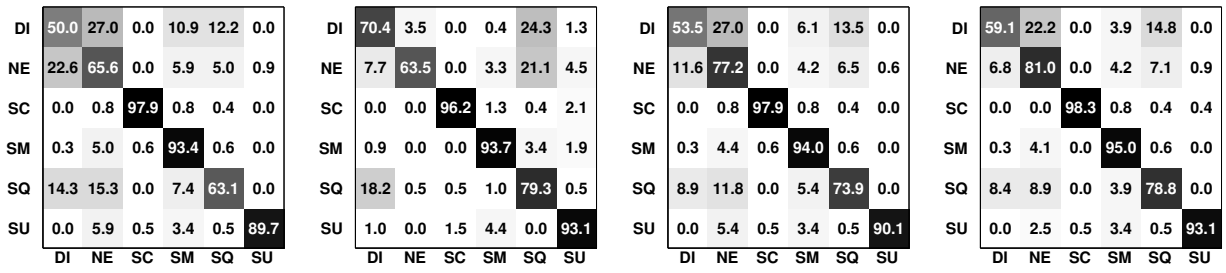
In this section, we evaluate the models in a multi-label classification scenario, where the adaptation is performed from multiple source domains. This is a more challenging setting, since the dataset is comprised of naturalistic facial expressions, and the recorded subjects are experiencing the affect in different ways and levels. The difficulty of the task can be seen in Table 2, where the subject-specific classifier GP_{target}, trained with 30 labeled data points, achieves a higher F1 score than the generic classifier GP_{source}, which is trained on 20 subjects. The adaptation attained by GPA and $GPDE_{ss}$ (the single source instance of GPDE) results in an improved average score compared to the subject specific GP_{target}. At this point note that GPA and $GPDE_{ss}$ perform similarly. The reason for this is that by treating all training subjects as a single source domain, $GPDE_{ss}$ smooths out the individual differences of the training subjects by treating them as data from a single, *broader*, source domain. Thus, the contribution of the target domain expert is diminished, as the variations of the target data can be explained, on average, by the source domain. On the contrary, the proposed GPDE with the adaptation from multiple sources (one per training subject) not only attains the best average F1 scores, but also achieves a more robust performance as evidenced from the higher AUC. Finally, note that with $N_t = 10$ GPDE performs better than the target specific classifier with $N_t = 500$. Note also that the proposed GPDE reaches the full (and the highest of all) performance with only 30 samples from the target domain. This is an important result, since obtaining the AU annotations (6 in this experiment) is expensive and time consuming.

Table 3 reports the detailed results (F1 score) per AU for the case of $N_t = 50$. The proposed GPDE attains a significant improvement (more than 5%) in AU4,6,25 compared to its counterparts, while it only suffers a loss from GPA on AU26. Moreover, the ROC curves in Fig. 4 show that GPDE exhibits a more robust performance not only on AUs with more pronounced improvement (*i.e.*, AU6), but also on AUs with similar F1 score to GPA (*i.e.*, AU12). The latter indicates that the proposed GPDE is a more robust model.

Table 3. F1 score for joint AU detection on DISFA. Subject adaptation with $N_t = 50$.

Method	AU4	AU6	AU9	AU12	AU25	AU26	Avg.
GP_{source}	51.93	42.34	41.06	58.89	78.84	57.98	53.17
GP_{target}	59.85	48.54	46.79	53.23	63.14	54.61	54.36
GPA [20]	56.75	47.97	43.88	**60.33**	78.35	**59.82**	57.85
$GPDE_{ss}$	60.20	50.90	**47.67**	59.17	73.82	59.17	58.49
GPDE	**65.59**	**53.62**	47.10	60.02	**79.96**	57.08	**60.56**

Finally, in Fig. 5 we demonstrate the ability of the proposed GPDE to fuse the predictions from the individual experts in order to form the overall prediction. In the selected example, we used the 20 first subjects from DISFA as the source domains, and correctly predicted the ground truth la-

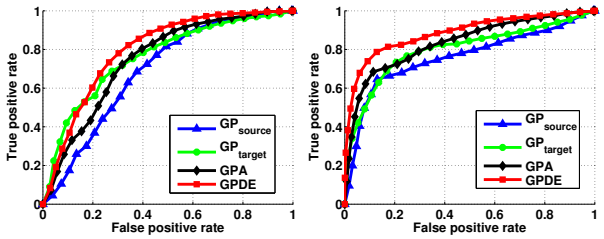


GP_{source} GP_{target} GPA GPDE

Figure 3. Confusion matrices averaged across the folds when using 50 target training data for $0° \rightarrow -30°$ adaptation.

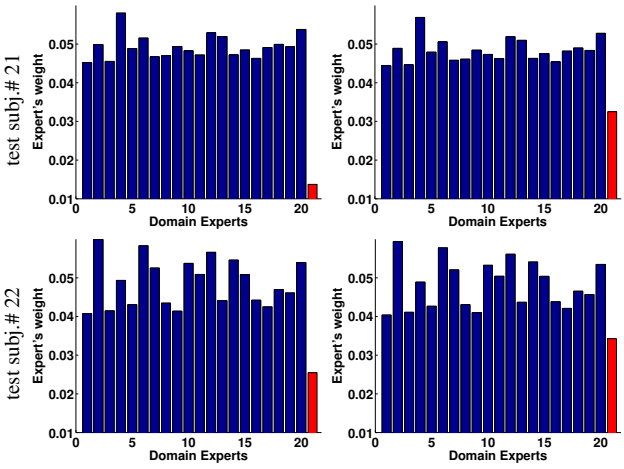
Table 2. Average results of 6 jointly predicted AUs with subject adaptation on DISFA.

Method	Average F1 score						Average AUC					
	10	30	50	100	200	500	10	30	50	100	200	500
GP_{source}	53.17						74.77					
GP_{target}	52.16	53.74	54.36	55.24	55.60	55.06	70.61	73.00	73.84	74.94	75.32	74.59
GPA [20]	56.54	57.42	57.85	57.87	58.22	58.39	76.45	77.64	78.14	78.52	79.07	79.38
$GPDE_{ss}$	56.27	57.74	58.49	58.76	59.12	58.88	75.04	77.83	78.72	79.23	79.67	79.08
GPDE	**58.66**	**60.04**	**60.56**	**60.18**	**60.48**	**60.17**	**78.25**	**80.15**	**80.59**	**80.20**	**80.27**	**79.86**



AU6 AU12

Figure 4. Average ROC curves for AU6 (left) and AU12 (right). Subject adaptation with $N_t = 50$.



$N_t = 10$ $N_t = 50$

Figure 5. Importance weight of each domain expert (20 source, 1 target) by means of normalized predicted precision for $N_t = 10$ (left) and $N_t = 50$ (right). The confidence of the target specific expert (red/last bar) increases as we increase the cardinality of the labeled target domain data. GPDE correctly predicts the activated AUs, *i.e.*, 12,25, in both cases.

bel (AU12 and AU25 active) for 2 different target subjects, *i.e.*, subj. #21 and #22. The depicted weights correspond to the normalized precisions of Eq. (13) and indicate a measure of confidence of each domain expert. The importance/confidence of the target expert increases when we use more labeled target data during the adaptation, as expected.

6. Conclusions

The work on domain adaptation in facial behavior analysis is still in its early stage. The conducted experiments on two adaptation tasks (view and subject) indicate several interesting facts: the source classifier trained on a large number of data can easily be outperformed by the classifier trained on as few as 50 examples from the target domain. Furthermore, the existing adaptation approaches try to adapt the target domain to the source domain by assuming that the two distributions can be matched. Yet, as we showed in our experiments on view adaptation, when more target data become available, the target classifier can largely outperform the existing adaptation approaches. The proposed model addresses these challenges by introducing the target expert, allowing it to reach (and outperform) the full performance of either source or target classifiers with as few as 30 target samples. In our future work, we plan to investigate the model adaptation to the other context factors (*i.e.*, 'when', 'why','what' and 'how'), and also to address modeling of the structure in the output (in the case of AU detection).

Acknowledgments

This work has been funded by the European Community Horizon 2020 under grant agreement no. 645094 (SEWA), and no. 688835 (DE-ENIGMA).

References

[1] T. Almaev, B. Martinez, and M. Valstar. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *ICCV*, pages 3774–3782, 2015.

[2] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[3] Y. Cao and D. J. Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014.

[4] J. Chen, X. Liu, P. Tu, and A. Aragones. Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*, 34(15):1964–1970, 2013.

[5] W.-S. Chu, F. D. L. Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR*, pages 3515–3522, 2013.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[7] M. P. Deisenroth and J. W. Ng. Distributed gaussian processes. In *ICML*, 2015.

[8] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell. Semi-supervised domain adaptation with instance constraints. In *CVPR*, pages 668–675, 2013.

[9] L. Duan, D. Xu, and I. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, 2012.

[10] P. Ekman, W. V. Friesen, and J. C. Hager. Facial action coding system. *Salt Lake City, UT: A Human Face*, 2002.

[11] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE TIP*, 24(1):189–204, 2015.

[12] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. D. la Torre. How much training data for facial action unit detection? In *FG*, 2015.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[14] N. Hesse, T. Gehrig, H. Gao, and H. K. Ekenel. Multi-view facial expression recognition using local appearance features. In *ICPR*, pages 3533–3536, 2012.

[15] J. Hoffman, B. Kulis, T. Darrell, and K. Saenko. Discovering latent domains for multisource domain adaptation. In *ECCV*, pages 702–715, 2012.

[16] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *ICLR*, 2013.

[17] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T. Huang. A study of non-frontal-view facial expressions recognition. In *ICPR*, pages 1–4, 2008.

[18] M. Kandemir. Asymmetric transfer learning with deep gaussian processes. In *ICML*, 2015.

[19] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, pages 1785–1792, 2011.

[20] B. Liu and N. Vasconcelos. Bayesian model adaptation for crowd counts. In *ICCV*, pages 4175–4183, 2015.

[21] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE TAC*, 4(2):151–160, 2013.

[22] Y.-Q. Miao, R. Araujo, and M. S. Kamel. Cross-domain facial expression recognition using supervised kernel mean matching. In *ICMLA*, pages 326–332, 2012.

[23] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011.

[24] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.

[25] C. Rasmussen and C. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

[26] O. Rudovic, M. Pantic, and I. Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE TPAMI*, 35(6):1357–1369, 2013.

[27] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE TPAMI*, 37(5):944–958, 2015.

[28] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR-W'*, 2013.

[29] E. Sangineto, G. Zen, E. Ricci, and N. Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *ACM Multimedia*, pages 357–366, 2014.

[30] M. Seeger. *Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations*. PhD thesis, University of Edinburgh, 2003.

[31] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *FG*, pages 921–926, 2011.

[32] T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *CVPR*, pages 2142–2150, 2015.

[33] G. Zen, E. Sangineto, E. Ricci, and N. Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In *ICMI*, pages 128–135, 2014.

[34] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE TPAMI*, 31(1):39–58, 2009.

[35] Z. Zhu and Q. Ji. Robust real-time face pose and facial expression recovery. In *CVPR*, volume 1, pages 681–688, 2006.