Coupled Gaussian Processes for Pose-Invariant Facial Expression Recognition

Ognjen Rudovic, *Student Member*, *IEEE*, Maja Pantic, *Fellow*, *IEEE*, and
Ioannis (Yiannis) Patras, *Senior Member*, *IEEE*

Abstract—We propose a method for head-pose invariant facial expression recognition that is based on a set of characteristic facial points. To achieve head-pose invariance, we propose the Coupled Scaled Gaussian Process Regression (CSGPR) model for head-pose normalization. In this model, we first learn independently the mappings between the facial points in each pair of (discrete) nonfrontal poses and the frontal pose, and then perform their coupling in order to capture dependences between them. During inference, the outputs of the coupled functions from different poses are combined using a gating function, devised based on the head-pose estimation for the query points. The proposed model outperforms state-of-the-art regression-based approaches to head-pose normalization, 2D and 3D Point Distribution Models (PDMs), and Active Appearance Models (AAMs), especially in cases of unknown poses and imbalanced training data. To the best of our knowledge, the proposed method is the first one that is able to deal with expressive faces in the range from $-45°$ to $+45°$ pan rotation and $-30°$ to $+30°$ tilt rotation, and with continuous changes in head pose, despite the fact that training was conducted on a small set of discrete poses. We evaluate the proposed method on synthetic and real images depicting acted and spontaneously displayed facial expressions.

Index Terms—Multiview/pose-invariant facial expression/emotion recognition, head-pose estimation, Gaussian process regression

1 Introduction

Facial expression recognition has attracted significant attention because of its various applications in psychology, medicine, security, and computing (human-computer interaction, interactive games, computer-based learning, entertainment, etc.) [1], [2]. Most of the existing methods deal with images (or image sequences) in which people depicted are relatively still and exhibit posed expressions in a nearly frontal view [3]. However, many real-world applications relate to spontaneous human-to-human interactions (e.g., meeting summarization, political debates analysis, etc.) where the assumption of having immovable subjects is unrealistic. This calls for a joint analysis of head-pose and facial expressions. Nonetheless, this remains a significant research challenge, mainly due to the large variation in appearance of facial expressions in different poses and difficulty in decoupling these two sources of variation.

Achieving accurate decoupling of rigid head motions from nonrigid facial motions so that the latter can be analyzed independently is the crux of any method for pose-invariant facial expression recognition. Previous research on pose-invariant facial expression recognition has addressed this problem either by assuming that rigid motions are independent of nonrigid facial motions (and therefore can be estimated sequentially and separately) (e.g., [4], [5]), or by simultaneously recovering these two sources of variation (e.g., [6], [7], [8]). Furthermore, the existing methods can be divided into face-shape-free methods (e.g., [9], [10], [11]) and face-shape-based methods (e.g., [12], [7], [13]). Face-shape-free methods achieve head-pose-invariance by using pose-invariant expression-related facial features extracted from 2D images, or by training the facial expression recognition method pose-wise. However, finding expression-related facial features independent of head pose is by no means an easy task because the changes in head-pose and facial expressions are nonlinearly coupled in 2D [7]. On the other hand, pose-wise facial expression recognition requires a large amount of training data in terms of different expressions and poses, which are often not readily available. In addition, the performance of the latter approach is expected to diminish when tested on facial images with continuous change in head pose. Face-shape-based methods rely on 2D/3D face-shape models that are used to decouple image variations caused by changes in facial expressions and head pose. However, since these methods are highly dependent on how well the shape models are aligned with the image data, which is not straightforward, the problem of facial expression recognition is inevitably compounded by the accuracy of the alignment [13].

In this work, we propose a probabilistic approach to head-pose-invariant facial expression recognition that is based on 2D geometric features, i.e., the locations of 39 characteristic facial points (see Fig. 1), extracted from an image depicting a facial expression of a subject with an arbitrary head pose. The output of the proposed method is the classified image, where the classification is performed in terms of six basic emotions (joy, anger, fear, disgust, surprise, and sadness) and neutral, proposed by Ekman

- *O. Rudovic and M. Pantic are with the Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom. E-mail: {o.rudovic, m.pantic}@imperial.ac.uk.*
- *I. Patras is with Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom. E-mail: i.patras@eecs.qmul.ac.uk.*

Manuscript received 29 June 2011; revised 15 May 2012; accepted 2 Oct. 2012; published online 25 Oct. 2012.
Recommended for acceptance by C. Sminchisescu.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-06-0417.
Digital Object Identifier no. 10.1109/TPAMI.2012.233.

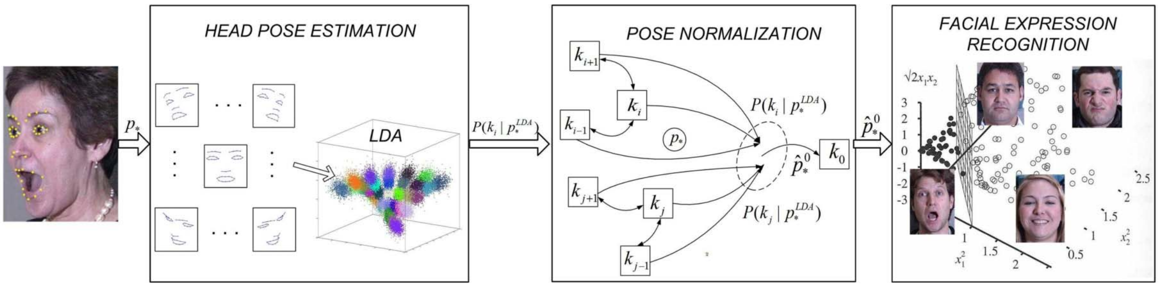


Fig. 1. The overview of the proposed approach. p_* are the 2D locations of the facial points extracted from the input face image, $P(k_i|p_*^{LDA})$ is the likelihood of p_* being in pose k_i, where k_0 is the frontal pose. The bidirectional lines in the pose normalization step connect the coupled poses, while the directed lines connect the poses for which the base SGPR models are learned. $\hat{p}_*^0$ is the prediction in the frontal pose for the query point p_*, obtained as a combination of the predictions obtained by the CSGPR models. The gating function is derived from the pose likelihoods $P(k_i|p_*^{LDA})$. Facial expression recognition is performed by applying a multiclass Support Vector Machine (SVM) classifier in the frontal pose to $\hat{p}_*^0$.

and Friesen [14]. The outline of the proposed method is given in Fig. 1, and it contains the following three steps: 1) head-pose estimation, 2) head-pose normalization, and 3) facial expression classification.

To perform the head-pose estimation, we first project the input facial points onto a low-dimensional manifold obtained by multiclass Linear Discriminant Analysis (LDA) [15]. We then use a Gaussian Mixture Model (GMM) [15], trained on the manifold data, to estimate the likelihood of the input being in a certain pose. In the second step, we perform the head-pose normalization. This is achieved by learning the mappings between a discrete set of nonfrontal poses and the frontal pose by means of the proposed Coupled Scaled Gaussian Process Regression (CSGPR) model. This model is built upon the Scaled Gaussian Process Regression (SGPR) model, used for learning the base mappings (with multiple outputs) between target pairs of poses (i.e., nonfrontal poses and the frontal pose). We propose the SGPR model for this task since it provides not only point predictions but also their uncertainty. The latter is explored in the CSGPR model to induce correlations between the base mappings, which are quantified in a form of a prior over the predictions of the base mappings and then incorporated into the base model, giving rise to a more robust regression model for head-pose normalization. To enable accurate head-pose normalization for continuous change in head pose (i.e., for poses that do not belong to the discrete set of poses), we devise a gating function that combines the point predictions made by the CSGPR models trained in discrete poses, and which is based on the head-pose estimation attained in the first step of the proposed approach. In the final step, we perform facial expression recognition by applying a multiclass Support Vector Machine classifier to the pose-normalized facial points.

The contribution of this work can be summarized as follows:

1. We propose a novel approach to head-pose-invariant facial expression recognition that can deal with expressive faces with head poses within the range from -45° to $+45^\circ$ pan rotation and -30° to $+30^\circ$ tilt rotation. The proposed approach performs accurately for continuous change in head pose, despite the fact that the training is conducted only on a small set of discrete poses.
2. We propose a novel probabilistic regression model for head-pose normalization, called Coupled Scaled Gaussian Process Regression. During inference, the CSGPR model selectively employs knowledge available in different poses, which results in a more accurate head-pose normalization than that achieved by the base mappings learned using state-of-the-art regression models and that achieved by the baseline methods such as 2D/3D Point Distribution Models (PDMs) and Active Appearance Models (AAMs).
3. The proposed approach to head-pose-invariant facial expression recognition can perform recognition of facial expression categories that were not present in certain nonfrontal poses during training. The existing head-pose-invariant facial expression recognition methods are trained pose-wise and therefore cannot deal with this scenario.

The rest of the paper is organized as follows: Section 2 gives an overview of the related work. Section 3 describes the proposed method for head-pose-invariant facial expression recognition. Section 4 describes the proposed CSGPR model for head-pose normalization. Section 5 presents the experimental results. Section 6 concludes the paper.

2 Head-Pose-Invariant Facial Expression Recognition: Related Work

Recent advances toward automatic head-pose-invariant facial expression recognition can be classified into face-shape-based approaches and face-shape-free approaches. Face-shape-based approaches use either 2D or 3D face-shape models to decouple rigid movements due to the change in head pose and nonrigid movements due to the changes in facial expressions. The methods proposed in [16], [17], and [18], for example, use 3D facial geometry deformation to recognize facial expressions in 3D images. These methods require high-quality capture of the facial texture and 3D geometrical data, and thus are not vastly applicable due to extensive and complex hardware requirements.

The methods in [5] and [19], for example, have an assumption that the 3D head pose is independent of the nonrigid facial movements. They use DBNs to model the unfolding of head-pose and nonrigid facial motions separately. Specifically, these methods first apply facial feature tracking and then estimate the 3D pose from the tracked facial points. The facial expression recognition is performed using pose-normalized facial features. The methods proposed in [20], [7], [13], and [8] try to decouple rigid head motions and nonrigid facial muscular motions simultaneously, by using 3D face models. For example, Kumano et al. [13] used a rigid 3D face-shape-model to extract person-specific facial descriptors from the face-shape model, which are then used in a particle filter framework to simultaneously estimate the head-pose and facial expressions. Several authors proposed Active Appearance Models for head-pose-invariant facial expression recognition (e.g., [21], [22], [23]). For example, Dornaika and Orozco [23] proposed an online AAM built upon a hierarchy of three AAMs (for eyebrows, lips, and eyelids and irises), which are used to estimate the 3D head pose and locations of characteristic facial points.

Although the methods mentioned above can be used to decouple the rigid and nonrigid facial motions, they require accurate alignment of the face-shape with the image data, which is challenging under varying facial expressions. Also, large out-of-plane head rotations, accounting for highly nonlinear changes in the appearance of an expression, are difficult to handle [13]. All these factors can diminish the performance of the facial expression recognition methods [8]. Some of the methods mentioned above, such as those based on AAMs, have to be trained per person/facial expression/head pose separately, which makes them difficult to apply in real-world scenarios where unknown subjects/facial expressions are expected. Person-independent AAMs are unable to deal with large variations in facial shapes and expressions, especially in the case of naturalistic data where variations in pose, facial morphology, and expression are large [24].

In contrast to increasing interest in head-pose-invariant facial expression recognition based on the 2D/3D face-shape models, head-pose-invariant facial expression recognition based on 2D face-shape-free models have seldom been investigated. Most of the proposed 2D face-shape-free head-pose-invariant methods address the problem of (expressionless) face recognition but not the problem of facial expression recognition (e.g., [25]). To date, only a few works have analyzed 2D-face-shape-free pose-invariant facial expression recognition [9], [26], [27], [28], [10], [29]. These approaches can be further divided into geometric- and appearance-feature-based approaches. Geometric-feature-based approaches rely on facial features such as shape of the face components and/or locations of the facial salient points (e.g., corners of the mouth). Appearance-feature-based approaches rely on skin texture changes such as wrinkles, bulges, and furrows. Both approaches have advantages and disadvantages. Geometric facial features cannot be used to describe textural changes like bulges and furrows. Appearance-based methods, though, are sensitive to changes in illumination and individual differences.

A typical appearance-based method is that by Zheng et al. [10]. The authors divide a facial image into subregions, and then extract SIFT descriptors from each subregion in the image, which are used as the input to the facial expression classifier based on k-Nearest Neighbors (k-NN). Moore and Bowden [9] applied a two step approach to pose-invariant facial expression recognition: head-pose classification, and pose-wise facial expression recognition based on Local Binary Patterns (LBPs) used as texture descriptors. Zheng et al. [28] proposed a hybrid approach that combines an appearance- and geometric-feature-based approach. Specifically, they used sparse SIFT features, extracted around 83 facial points, as the input to the k-NN facial expression classifier. Apart from our previous work in [29], the only work based solely on geometric features is that proposed by Hu et al. [27], who investigated facial expression recognition in nonfrontal poses based on locations of 83 facial points. This work is based on a set of facial images with five yaw angles generated using the BU-3DFE dataset [18]. A shortcoming of this method and most of the aforementioned methods is that they perform pose-wise facial expression recognition. Consequently, the performance of these methods is expected to diminish when tested on data in the poses not used to train the classifiers (i.e., nondiscrete poses) [13]. Furthermore, these methods require a large amount of (annotated) facial data per pose in order to train the classifiers. More importantly, they cannot perform recognition of facial expressions that were not available in certain poses during training (in other words, they cannot deal with novel facial expression categories). Our work aims at addressing these limitations of the existing methods by means of the Gaussian process regression framework.

3 Head-Pose-Invariant Facial Expression Recognition

In this section, we describe the proposed approach to head-pose-invariant facial expression recognition. We use the locations of 39 characteristic facial points as the input, although a different number of the points can be used instead. The locations of the facial points can be extracted either manually or automatically. To date, several different methods for automatic facial point localization have been proposed (e.g., [30]). In this work, automatic localization of the facial points is achieved using the online AAM proposed in [23]. The pose-invariant facial expression recognition is then performed in three steps: 1) head-pose estimation, 2) head-pose normalization, and 3) facial expression classification in the frontal pose. These steps are described in the following sections and are summarized in Algorithm 1.

```
Algorithm 1. Head-pose-invariant Facial Expression Recognition
Input: Positions of facial landmarks in an unknown pose ($p_*$)
Output: Facial expression label ($l$)
1. Apply the head-pose estimation (Section 3.1) to obtain
   $P(k|p_*^{lda})$, $k = 0, \dots, P-1$.
2. Register $p_*$ to poses $k \in \mathcal{K}$ which satisfy $P(k|p_*^{lda}) > P_{min}$
   (Section 3.2), and predict the locations of the facial
   landmarks points in the frontal pose (Section 4) as
       $\hat{p}_*^0 = \frac{1}{\sum_{k \in \mathcal{K}} P(k|p_*^{lda})} \sum_{k \in \mathcal{K}} P(k|p_*^{lda}) f_C^{(k)}(p_*^k)$.
3. Facial expression classification in the frontal pose
   (Section 3.3)
       $l \leftarrow \arg\max_z (\sum_{i:p_i^0 \in T_z} \alpha_i^z K(p_i^0, \hat{p}_*^0) + b_z)$.
```

In what follows, we first divide the head-pose space into $P = 35$ evenly distributed discrete poses ranging from $-45°$ to $+45°$ pan rotation and $-30°$ to $+30°$ tilt rotation, with an increment of $15°$. The locations of d facial points extracted from an expressive face in pose k, where $k = 0, \dots, P-1$, are stored in a vector $p^k \in \mathbf{R}^{2d}$. The training dataset is then denoted by $D = \{D^0, \ldots, D^k, \ldots, D^{P-1}\}$, where $D^k = \{p_1^k, \ldots, p_N^k\}$ is comprised of N training examples in a nonfrontal pose k, with D^0 containing the corresponding training data in the frontal pose.

3.1 Head-Pose Estimation

Different methods for head-pose estimation based on appearance and/or geometric features have been proposed (see [31] for an overview). In this work, we devise a simple but efficient method for head-pose estimation that is based on multiclass LDA [15]. To this end, we first align the facial points in each discrete pose by using generalized Procrustes analysis to remove the effects of scaling and translation. Then, we learn a low-dimensional manifold of head poses by means of multiclass LDA [15] using the aligned training data in each discrete pose and the corresponding head-pose labels. This manifold encodes head-pose variations while ignoring other sources of variations such as facial expressions and intersubject variation. We denote the vector of the input facial points projected onto this manifold as p_{lda}. The distribution of such vectors having the same head pose is modeled using a single Gaussian. Consequently, the likelihood of a test input p^*_{lda} being in pose k is then given by $P(p^*_{lda}|k) = \mathcal{N}(p^*_{lda}|\mu_k, \Sigma_k)$, where μ_k and Σ_k are the mean and covariance of the training data in pose k after being projected onto the pose manifold. By applying Bayes' rule, we obtain $P(k|p^{lda}) \propto P(p^{lda}|k)P(k)$, where a uniform prior over the poses is used.

3.2 Head Pose Normalization

The head-pose normalization is attained by mapping the locations of the facial points from an arbitrary head pose to the locations of the corresponding facial points in the frontal pose. To this end, we apply the proposed CSGPR model, which is explained in detail in Section 4.

3.3 Facial Expression Classification in Frontal Pose

The final step in the proposed approach is the facial expression classification applied to the pose-normalized facial points. To this end, different classification methods can be employed (e.g., see [3], [32]). We adopt the multiclass SVM classifier [33], with the one-versus-all approach, since this classifier has commonly been used in the facial expression recognition tasks (see [9]). Briefly, the SVM classifier takes the locations of the facial points in the frontal pose $\hat{p}^0_*$ as the input and constructs a separating hyperplane that maximizes the margin between the positive and negative training examples for each class. Formally, the labels l for each expression class take $0/1$ value, and are obtained as

$$l = \arg\max_z \left(\sum_{i:p_i^0 \in T_z} \alpha_i^z K(p_i^0, \hat{p}^0_*) + b_z \right), \; z = 1, \ldots, Z, \tag{1}$$

where α_i^z and b_z are the weight and bias parameters, $K(p_i^0, \hat{p}^0_*)$ is a vector of the inner products between the training points $p_i^0 \in D^0$ and the predicted points $\hat{p}^0_*$. The set denoted by T_z contains training examples of the zth facial expression class.

3.4 Head-Pose-Invariant Facial Expression Recognition: Algorithm Summary

Algorithm 1 summarizes the proposed approach. Given a query point (p_*), we first compute the likelihood of its being in a nonfrontal pose k ($P(k|p^{lda}_*)$), where $k = 1, \ldots, P-1$. The facial points in the frontal pose ($\hat{p}^0_*$) are then obtained as a weighted combination of the predictions of the coupled functions $f_C^{(k)}(p_*)$ from nonfrontal poses. Note that before $f_C^{(k)}(\cdot)$ is applied to the points p_*, these points are first registered to a reference face in the pose k, which is a standard preprocessing step. This registration is performed by applying an affine transformation learned using three referential points: the nasal spine point and the inner corners of the eyes. These are chosen since they are stable facial points, and are not affected by facial expressions [32]. The facial expression classification is then performed by applying the multiclass SVM classifier to the pose-normalized facial points. Finally, note that the inference time for p_* can be significantly reduced by considering only the most likely poses, i.e., $P(k|p^{lda}_*) > P_{min}$, where P_{min} is chosen so that only the predictions from the poses being in the vicinity of the test input p_* are considered.

4 Coupled Scaled Gaussian Process Regression

In this section, we describe the proposed CSGPR model for head-pose normalization. For this, we first learn a set of base functions $\{f^{(1)}(\cdot), \ldots, f^{(k)}(\cdot), \ldots, f^{(P-1)}(\cdot)\}$ for mapping the facial points from nonfrontal poses to the corresponding points in the frontal pose. An ensemble of coupled function $\{f_C^{(1)}(\cdot), \ldots, f_C^{(k)}(\cdot), \ldots, f_C^{(P-1)}(\cdot)\}$ is then inferred by modeling the correlations between the base functions. In this way, we perform knowledge transfer across the poses. This is important when different training data, in terms of variety in facial expressions, are available in different poses because it may improve the performance of the base mappings learned per pose and independently from other poses.

4.1 Scaled Gaussian Process Regression

To learn the base mapping functions $f^{(k)}(\cdot)$, we propose Scaled GPR. This model is based on the Scaled Gaussian Process Latent Variable Model (SGPLVM) proposed in [34]. However, SGPLVM is designed for dimension reduction, a different problem from supervised learning we consider here. In contrast to standard GPR [35], which deals with a single output only (i.e., each coordinate of each facial point), SGPR is specifically designed for simultaneous prediction of multiple outputs (i.e., all coordinates of all facial points).

Formally, given a collection of N_k training pairs of the facial points in a nonfrontal pose k and the corresponding points in the frontal pose, $\{D^k, D^0\}$, where each element p_i^k and p_i^0 ($i = 1, \ldots, N_k$) in D^k and D^0 is a $2d$-dimensional vector (d is the number of the facial points), the goal is to learn the mapping

$$p^0 = f^{(k)}(p^k) + \mathbf{1}_{1\times(2d)}\varepsilon_i, \tag{2}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ and σ_n^2 is the noise variance.

For standard GPR with a single output dimension m, the likelihood of the output given the inputs is

$$p(\{p^0_{i,m}\}|\{p^k_{i,m}\},\theta)=\frac{1}{\sqrt{(2\pi)^{N_k}|K|}}\exp\Big(-\frac{1}{2}D^k_m K^{-1}\big(D^k_m\big)^T\Big). \tag{3}$$

SGPR introduces a scaling parameter w_m for each output dimension m, which is equivalent to defining a separate kernel function $k(x_i,x_j)/w^2_m$ for each output [34]. Plugging this into the GPR likelihood for $m=1,\dots,2d$, the complete joint likelihood of the SGPR model is

$$\begin{aligned}&p(\{p^0_i\}|\{p^k_i\},\theta,W)\\&=\prod_m\frac{w^2_m}{\sqrt{(2\pi)^{N_k}|K|}}\exp\Big(-\frac{1}{2}w^2_m D^k_m K^{-1}\big(D^k_m\big)^T\Big),\end{aligned} \tag{4}$$

where $\theta=\{\sigma_s,S,\sigma_l,\sigma_b,\sigma_n\}$ are the kernel parameters and $W=\{w_1,\dots,w_{2d}\}$. The covariance matrix K has the entries

$$\begin{aligned}k(p^k_i,p^k_j)&=\sigma^2_s\exp\Big(-\frac{1}{2}(p^k_i-p^k_j)^T S^{-1}(p^k_i-p^k_j)\Big)\\&\quad+\sigma_l p^k_i p^k_j+\sigma_b,\quad i,j=1,\dots,N_k,\end{aligned} \tag{5}$$

where σ^2_s is the variance and $S=diag(s^2_1,\dots,s^2_{2d})$ are the length-scales of each input dimension (i.e., each coordinate of each landmark point) of the RBF kernel, σ_l is the scale of the linear kernel, and σ_b is the model bias. We adopt this composite kernel because it can handle both linear and nonlinear data structures [35].

The model parameters θ and W are found by minimizing the negative log-likelihood

$$\begin{aligned}&-\ln p(D^k,\theta,W|D^0)=d\ln|K|\\&+\frac{1}{2}\sum_{m=1}^{2d}w^2_m\big(D^0_m\big)^T(K+\sigma^2_n I)^{-1}D^0_m+const.\end{aligned} \tag{6}$$

This likelihood function is first minimized w.r.t. θ using Scaled Conjugate Gradient algorithm [35]. The scale parameters W are then computed in the closed form as $w_m=\sqrt{2d/((D^0_m)^T K^{-1}D^0_m)}$. These two steps are repeated until convergence of the likelihood function.

During inference in SGPR, the mean $f^{(k)}(p^k_*)$ and variance $V^{(k)}(p^k_*)$ of the predictive distribution for the query point p^k_* are obtained as

$$f^{(k)}(p^k_*)=k^T_*(K+\sigma^2_n I)^{-1}D^0, \tag{7}$$

$$V^{(k)}\big(p^k_*\big)=\big(k\big(p^k_*,p^k_*\big)-k^T_*(K+\sigma^2_n I)^{-1}k_*\big)\,diag(W)^{-2}, \tag{8}$$

where $k_*=k(D^k,p^k_*)$. The mean $f^{(k)}(p^k_*)$ provides point predictions of the facial points in the frontal pose, and $V^{(k)}(p^k_*)$ their uncertainty .

4.2 Learning Coupled Functions

So far, we have used SGPR to learn a set of the base functions that map the facial points from nonfrontal poses to the frontal pose. However, since these functions are learned separately, there is no sharing of knowledge among poses. This sharing may be valuable when different training data are available across the poses. We accomplish this sharing by learning a set of coupled functions which take into account the correlations between the base mappings. This is illustrated by an example of coupling a function $f^{(k_2)}(\cdot)$, the base function for pose k_2, to a function $f^{(k_1)}(\cdot)$, the base function for pose k_1. We adopt a parametric approach to learning the correlations between the mapping functions, which are induced through a prior distribution defined as

$$P(f^{(k_1)},f^{(k_2)}|k_1)\propto\exp\Bigg(-\frac{1}{2\sigma^2_{(k_1,k_2)}}\big(f^{(k_1)}\big(p^{k_1}_*\big)-f^{(k_2)}\big(p^{k_1}_*\big)\big)\Bigg), \tag{9}$$

where $\sigma^2_{(k_1,k_2)}$ is the variance of coupling that is estimated from training data D^{k_1} and D^{k_2}. Intuitively, it measures the similarity of the predictions made by the function $f^{(k_2)}(\cdot)$ and predictions made by the function $f^{(k_1)}(\cdot)$, when they are evaluated on the training data in pose k_1. It can also be seen as an independent noise component in the predictions obtained by $f^{(k_2)}(\cdot)$, which is learned using training data in pose k_2, when evaluated on training data in pose k_1. Because we assume that this noise is Gaussian and independent of the noise already modeled in $f^{(k_2)}(\cdot)$, these two sources of randomness simply add [35]. Consequently, by including the coupling variance $\sigma^2_{(k_1,k_2)}$ into predictive distribution of $f^{(k_2)}(\cdot)$, we obtain the following expressions for the mean and variance of the predictive distribution of the coupled function $f^{(k_1,k_2)}(\cdot)$ as

$$f^{(k_1,k_2)}(p^{k_1}_*)=k^T_{k_2*}(K_{k_2}+\big(\sigma^2_{nk_2}+\sigma^2_{(k_1,k_2)}\big)I)^{-1}D^0, \tag{10}$$

$$\begin{aligned}V^{(k_1,k_2)}(p^{k_1}_*)&=(k_{k_2}\big(p^{k_1}_*,p^{k_1}_*\big)\\&\quad-k^T_{k_2*}(K_{k_2}+\big(\sigma^2_{nk_2}+\sigma^2_{(k_1,k_2)}\big)I)^{-1}k_{k_2*})\,diag(W_{k_2})^{-2},\end{aligned} \tag{11}$$

where the subindex k_2 refers to the model parameters of the base function for pose k_2, and $k_{k_2*}=k(D^{k_2},p^{k_1}_*)$. Here, the sharing of knowledge between the poses k_1 and k_2 is achieved through the coupled function $f^{(k_1,k_2)}(\cdot)$, which uses training data from pose k_2 when making predictions from pose k_1. Note also from (11) that the less $f^{k_2}(\cdot)$ is coupled to $f^{k_1}(\cdot)$, which is measured by the coupling variance $\sigma^2_{(k_1,k_2)}$, the higher the uncertainty in the outputs obtained by the coupled function $f^{(k_1,k_2)}$. In other words, if the functions are perfectly coupled (i.e., $\sigma^2_{(k_1,k_2)}\to 0$), then $f^{(k_1,k_2)}(\cdot)\to f^{(k_2)}(\cdot)$. Conversely, if they are very different (i.e., $\sigma^2_{k_1,k_2}\to\infty$), then $f^{(k_1,k_2)}(\cdot)$ converges to a GP prior with the zero mean and constant variance. Last, the variance in (11) is guaranteed to be positive definite since we add a positive term (i.e., the coupling variance) to its diagonal.

4.3 CSGPR: Model

In the previous section, we introduced the concept of the coupled functions. In this section, we explain how the outputs of the base and coupled functions are combined, resulting in the Coupled SGPR model for head-pose normalization. Let us consider the base function $f^{(k_2)}(\cdot)$ and the coupled function $f^{(k_1,k_2)}(\cdot)$. During inference, these two functions give their own predictions of the facial points in the frontal pose. We now combine them in order to obtain a single prediction. A straightforward approach is to apply either Density-Based (DB) weighting, using the pose estimation explained in Algorithm 1, or the Variance-Based (VB) weighting, where the weights are set to inversely proportional values of the uncertainty in GP predictions. In this work, we employ the Covariance Intersection (CI) [36] rule for combining predictions, which is the optimal fusion rule when correlation between the prediction errors of two estimators are unknown [37]. For predictions obtained by the base and coupled functions, this fusion rule yields the mean and the variance of the CSGPR model, given by

$$\begin{aligned} f_C^{(k_1)}(p_*) = V_C^{k_1}(p_*)(\omega V^{(k_1)}(p_*)^{-1} f^{(k_1)}(p_*) \\ + (1-\omega) V^{(k_1,k_2)}(p_*)^{-1} f^{(k_1,k_2)}(p_*)), \end{aligned} \tag{12}$$

$$V_C^{(k_1)}(p_*)^{-1} = \omega V^{(k_1)}(p_*)^{-1} + (1-\omega) V^{(k_1,k_2)}(p_*)^{-1}. \tag{13}$$

The optimal $\omega \in [0,1]$ is found during inference by minimizing the trace of $V_C^{(k_1)}(p_*)$, used as the uncertainty criterion, w.r.t. ω (see [36] for details).

4.4 Pruning Scheme

Making inference with all coupled functions, i.e., $\frac{P(P-1)}{2}$ coupled functions, would be computationally intensive. Also, not all the coupled functions contribute to improving the predictions obtained by the base functions. For this, we propose a pruning criterion based on the number of Effective Degrees of Freedom (EDoF) [38] of a GP to select the coupled functions that will be used during inference. EDoF of a GP measures how many degrees of freedom are used by the given data, and can be a good indicator of the variability in the training dataset (in terms of facial expressions). Hence, in our pruning scheme we keep only the coupled functions that have a similar or larger number of EDoF than that of the base functions they are coupled to. In this way, we significantly reduce the computational load of the CSGPR model during inference. We define the number of EDoF of a coupled function $f^{(k_1,k_2)}(\cdot)$ as

$$C_{eff}^{(k_1,k_2)} = \sum_{i=1}^{N_{k_2}} \frac{\lambda_{k_2}^i}{\lambda_{k_2}^i + \sigma_{nk_2}^2 + \sigma_{(k_1,k_2)}^2}, \tag{14}$$

where $\lambda_{k_2}^i$ are the eigenvalues of the covariance matrix K_{k_2} and N_{k_2} is the number of training data used to learn the base function $f^{k_2}(\cdot)$. The number of EDoF is approximately equal to the number of eigenvalues of the kernel matrix K_{k_2} that are greater than the noise variance. Thus, if $\sigma_{(k_1,k_2)}^2$ is high, then $C_{eff}^{(k_1,k_2)} \to 0$, and the predictions made by the coupled function $f^{(k_1,k_2)}(\cdot)$ can be ignored. The coupled functions used for inference are selected based on the ratio: $C_{eff}^{(k_1,k_2)}/C_{eff}^{(k_1)}$, where its minimum value (C_{min}) is set using a cross-validation procedure, as explained in the experiments. The number of EDoF of the base functions is computed using (14) without the coupling variance term. Note that the coupling variance could also be used as a criterion for pruning. However, the proposed measure is more general since it also tells us how much we can "rely" on the coupled function in the presence of novel data (e.g., novel facial expression categories)—something that is not encoded by the coupling variance. Finally, learning and inference in CSGPR are summarized in Algorithm 2.

```
Algorithm 2. Learning and inference with CSGPR
OFFLINE: Learn the base SGPR models and coupling
            variances
1. Learn $P-1$ base SGPR models $\{f^{(1)}(\cdot), \ldots, f^{(P-1)}(\cdot)\}$ for
   target pairs of poses (Section 4.1).
2. Perform coupling of the base SGPR models learned in
   Step 1
     for $k_1 = 1$ to P − 1 do
       for $k_2 = 1$ to P − 1 and $k_1 \neq k_2$ do
         predict $\sigma_{(k_1,k_2)}$ (Section 4.2)

           if $C_{eff}^{(k_1,k_2)}/C_{eff}^{(k_1,k_1)} > C_{min}$ then

             $\sigma_C^{k_1} = [\sigma_C^{k_1}, \sigma_{(k_1,k_2)}]$
           end if
       end for
       store $\sigma_C^{k_1}$
     end for
ONLINE: Infer the facial points in the frontal pose from the
           facial points $p_*^{k_1}$ in pose $k_1$
$B_{k_1}$: number of the base functions coupled to $f^{(k_1)}$
1. Evaluate the base function for pose $k_1$ (Section 4.1):
   $Pr(0) = \{f^{(k_1)}(p_*^{k_1}), V^{(k_1)}(p_*^{k_1})\}$.
2. Combine the functions coupled to pose $k_1$ (Section 4.2)
   for $i = 1$ to $B_{k_1}$ do
     $\sigma_{(k_1,i)} = \sigma_C^{k_1}(i)$ , $Pr(i^-) = \{f^{(k_1,i)}(p_*^{k_1}), V^{(k_1,i)}(p_*^{k_1})\}$
     $Pr(i^+) = CI(Pr(i-1), Pr(i^-))$ (Section 4.3)
   end for
     $\{f_C^{(k_1)}(p_*^{k_1}), V_C^{(k_1)}(p_*^{k_1})\} = Pr(i)$.
```

4.5 Multi-Output GPR: Related Work

In this section, we give a brief overview of GPR models that can also be used to learn the base mappings. Standard GPR deals with a single output and cannot be used to *jointly* map the locations of the facial points from nonfrontal poses to the frontal pose. Modeling of each coordinate of each facial point independently is possible, but the learned mappings will be suboptimal because the interactions between the points are ignored.

Recent research on GPR has focused on learning the interactions between the output dimensions, e.g., [39], [40], [41], [42], [43]. Boyle and Frean [39] induce correlations between two outputs by deriving the output processes as different convolutions of the same underlying white noise process. A generalization to more than two outputs has been proposed by Alvarez and Lawrence [42]. Yu et al. [40] proposed to share models among the outputs by learning separate GPs for each output, but assuming that their parameters are drawn from the same prior. Bonilla et al. [41] proposed a model that learns a shared covariance function on input-dependent features and a "free-form" covariance matrix over outputs. A different approach to modeling multiple outputs is proposed by Bo and Sminchisescu [43], where a GP prior is placed over the outputs, and the inference is carried out by minimizing the Kullback-Leibler (KL) divergence between the input and the output GPs, modeled as normal distributions over training and testing examples. These models have been empirically shown to outperform the GPR models trained independently for each output. However, this holds only if

TABLE 1
Summary of the Used Data from the Employed Datasets

Dataset	Subjects	Expressions		Poses			Type	
		number	levels	tilt	pan	total	acted	real
BU3DFE	100	7	2	(-30°,+30°)	(-45°,+45°)	247	√	×
MulitPIE	50	4	1	0°	(-45°,0°)	4	√	√
MPFE	3	7	1	(-30°,+30°)	(-45°,+45°)	∞	√	√
Semaine	10	2	∞	(-30°,+30°)	(-45°,+45°)	∞	×	√

We use ∞ for facial expression levels and poses that change continuously.

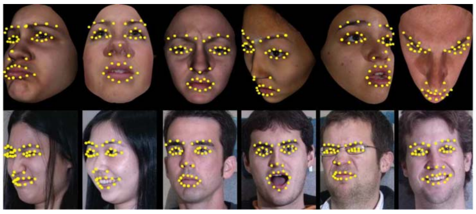


Fig. 2. Example images from the **BU3DFE** dataset (*top*) and the **MultiPIE** dataset (*bottom*) with synthetic and manually localized facial points, respectively.

the sharing of knowledge among the outputs exists, i.e., when the outputs are correlated and some of them are endowed with more training data than the others; otherwise, there could be no benefit in using these models [41].

In the proposed SGPR for learning the base functions with multiple outputs, the correlations among different outputs are induced through the shared covariance function and scaling parameters defined for each output. Note that modeling the couplings *between* the base functions with multiple outputs, which are learned per different pairs of poses, is *different* from modeling dependences *within* the outputs (i.e., facial points) of a single function with multiple outputs, learned per pose. The latter task can be attained by the models in [39], [42], [40], [42] and the SGPR model, in which covariance can be seen as a scaled (per output) covariance of the model in [40], and a block-diagonal counterpart of the full covariance used in [41]. Nevertheless, the inference time of the SGPR model is equal to that of the GPR model with a single output, i.e., $O(N^2)$, where N is the number of training examples. This is considerably lower compared to the inference time of most of the GPR models mentioned above for multiple outputs (which scales as $O(MN^2)$, where M is the number of outputs)—especially when M is large, as in our case $M = 78$. More importantly, these models are not directly applicable to the task of modeling dependencies *between* the base functions with multiple outputs learned per pose since they assume the same input features for predicting a multidimensional output. In other words, these methods use the same covariance function for all outputs, which makes them unfit for the target task. Finally, the Bayesian Co-training [44] framework for GP multiview learning has recently been proposed. However, this framework is not directly suited for the target task since it cannot deal with problems where the evidence from different views should be additive (or enhanced), as in our case, rather than averaging [44].

5 Experiments

5.1 Datasets and Experimental Procedure

To evaluate the proposed method, we used facial images from three publicly available datasets: the BU-3D Facial Expression (BU3DFE) [18] dataset, the CMU Pose, Illumination and Expression (MultiPie) [45] dataset, and the Semaine [46] dataset. We also used the Multipose Facial Expression (MPFE) dataset that we recorded in our lab. Table 1 summarizes the properties of each dataset, and Figs. 2 and 7 show the sample images from the datasets. The BU3DFE and the MPFE datasets contain images depicting facial expressions of Anger (AN), Surprise (SU), Disgust (DI), Joy (JO), Sadness (SA), Fear (FE), and Neutral (NE). From the MultiPIE dataset, we used images of facial expressions of SU, DI, JO, and NE, and from the Semaine dataset we used 10 image sequences, coded per frame either as Speech or Laughter. The facial expressions in the BU3DFE dataset are acted at four different levels of intensity, where the highest level corresponds to the apex of the expression. The facial expressions in the MultiPIE and the MPFE datasets are also acted and depict only the apex of the expressions, while the images in the Semaine dataset are spontaneously displayed. In the case of the BU3DFE dataset, we rendered 2D facial images of 100 subjects (58 percent female) at *levels* 3 and 4 of the expression, and in 247 discrete poses (with 5 percent increment in pan and tilt angles), using the 3D range data. Images from all 247 poses were used during testing, whereas images from a subset of 35 poses (with 15 percent increment in pan and tilt angles) were used for training. The images from the MultiPie dataset depict 50 subjects (22 percent female) captured at 4 pan angles ($0°$, $-15°, -30°$, and $-45°$). The MPFE dataset contains expressive images of three subjects (33 percent female) and the Semaine dataset contains expressive images of 10 subjects (60 percent female), with various head poses. All the images were annotated in terms of 39 facial points (e.g., see Fig. 2). Specifically, the MultiPIE dataset was annotated manually, while for the BU3DFE dataset the locations of the facial points are provided by the dataset creators. The facial images from the MPFE and the Semaine dataset were annotated automatically using the AAM tracker [23].

The training dataset contained the locations of the facial points in 34 nonfrontal poses, and the corresponding facial points in the frontal pose (thus, 35 poses in total). The training points were registered per pose, as explained in Section 3.4. For testing, we used the facial points from the training poses (tp) and the nontraining poses (ntp). We measured the performance of the head-pose normalization using the Root Mean of the Squared Error (RMSE) computed between the pose-normalized facial points and the ground truth in the frontal pose. The performance of the facial expression recognition was measured using the Recognition Rate (RR) computed by applying the SVM classifier (F-SVM), trained using training data in the frontal pose, to the pose-normalized facial points. If not stated otherwise, we applied fivefold cross validation in all our experiments, with each fold containing images of different subjects.

We compared the performance of the proposed regression model for head-pose-normalization to Linear Regression (LR) and Support Vector Regression (SVR)[47]. We performed further comparisons of the proposed model with

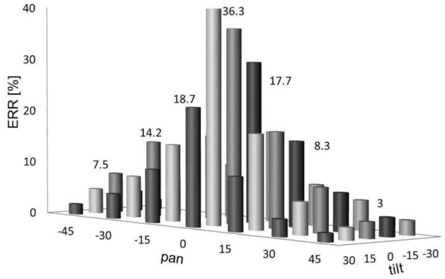


Fig. 3. The error rate (ERR) per pose attained by the LDA-based head-pose classification. The subspace of head poses was learned using $N = 200$ training data-pairs from each of the 35 training poses from the **BU3DFE** dataset. The average ERR is 9 percent.

recently proposed models for multi-output GPR: Twin GPR (TWINGPR)[43] and Multitask GPR (MTGPR) [41]. As a baseline for these methods, we used Independent GPRs (IGPRs)[35] for each output (i.e., coordinate of each facial point). Also, analogously to the coupling of the SGPR models, we performed coupling of the IGPR models to obtain the Coupled IGPR (CIGPR) models. We did so since IGPR has the same covariance form as SGPR and thus the coupling of the IGPR models using the proposed framework is straightforward. Apart from TWINGPR, the hyperparameters of all other GPR-based models were optimized by minimizing the negative log-likelihood of the models. In the case of TWINGPR, SVR, and the pose-wise SVMs (PW-SVMs), we cross-validated the model parameters. In all models, we used a composite kernel function that is a sum of a linear term, an isotropic Radial Basis Function (RBF), and a model bias. We also include the results obtained by the 2D-PDM [48] and 3D-PDM [7], and the AAM [23].

5.2 Experiments on Synthetic Data

In this section, we present the experiments conducted on the BU3DFE dataset. Fig. 3 shows the error rate for head-pose classification attained by taking the most likely discrete head pose as the predicted class. The likelihood of each head pose was obtained by the head-pose estimation approach described in Section 3.2. As can be seen, the larger misclassification occurs in near-frontal poses. This is to be expected since the facial points in near-frontal poses are more alike than those in nonfrontal poses that are far from the frontal pose. Note also that the misclassification occurs mostly among the neighboring poses, which is a tractable problem for the CSGPR model due to the definition of its weighting function (see Algorithm 1, step 2).

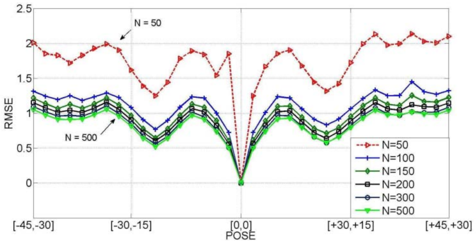


Fig. 4. RMSE of head-pose normalization attained by the CSGPR model trained per pose and by using N data-pairs from the **BU3DFE** dataset.

We next compare the performance of different models w.r.t. the number of data used for training (when the head pose is known). To train the models, we used N training data-pairs per each pair of a nonfrontal pose and the frontal pose, sampled uniformly from all the expression classes at random and from four folds. The fifth fold was used to test such trained models. This was repeated for all the folds, and the average RMSE for head-pose normalization, attained by different regression models, w.r.t. the number of training data-pairs N is shown in Table 2. We also include the recognition results attained by the PW-SVM classifiers. Note that MTGPR, specifically designed for dealing with multiple outputs, fails to outperform the other GP-based regression models in the target task. We noticed from the training and testing performance of this model that, for the given range of N, it was prone to overfitting. This is probably because of the large number of the model outputs, resulting in the large number of MTGPR parameters to be learned. On the other hand, TWINGPR performs better pose-normalization (in terms of RMSE). However, this does not translate into the RR attained by this model, compared to that of IGPR and SGPR and their coupled counterparts, which outperform the other tested models in the targets task. Finally, note that the PW-SVM classifiers require more training data to achieve the RR similar to that of the GPR-based methods, yet it remains lower than that attained by the coupled models. Fig. 4 shows the performance of CSGPR-based head-pose normalization across different discrete poses w.r.t. the number of training data N. It can be noted that this model exhibits stable performance across the poses. The experiments show evidence that the coupled models generalize well (and better than the other tested

TABLE 2
RMSE and RR Attained by the Base Models for Head-Pose Normalization and Facial Expression Recognition

Method	RR (%)						RMSE (in pixel)					
	N=50	N=100	N=150	N=200	N=300	N=500	N=50	N=100	N=150	N=200	N=300	N=500
LR&F-SVM	57.2	59.8	62.3	67.0	67.9	68.4	2.71	2.45	2.11	1.72	1.80	1.83
SVR&F-SVM	64.3	68.4	70.3	71.5	71.9	72.1	1.85	1.42	1.27	1.18	1.15	1.09
TWINGPR&F-SVM	64.0	69.2	70.7	71.4	71.8	72.3	2.18	1.35	1.15	0.90	0.85	**0.80**
MTGPR&F-SVM	61.6	65.4	67.8	69.1	69.8	70.1	2.11	1.74	1.43	1.35	1.28	1.26
IGPR&F-SVM	68.1	70.6	72.0	72.3	72.8	73.2	1.71	1.22	0.98	0.93	0.88	0.83
SGPR&F-SVM	66.8	70.5	71.8	72.1	72.4	72.9	1.72	**1.15**	1.01	0.95	0.89	0.85
CIGPR&F-SVM	**69.2**	**72.2**	**73.1**	73.7	74.0	74.6	**1.68**	1.19	**0.95**	0.92	0.85	0.81
CSGPR&F-SVM	68.7	72.1	72.9	**73.9**	**74.2**	**74.9**	1.70	1.17	0.98	**0.90**	**0.82**	0.82
PW-SVM	60.3	66.4	68.5	70.4	72.7	73.3	-	-	-	-	-	-

The models were trained using N data-pairs per pose from the ***BU3DFE*** *dataset. In the case of regression-based methods, the classification was performed by applying F-SVM classifier to the pose-normalized facial points.*

TABLE 3
The Performance of Different Methods for Head-Pose-Invariant Facial Expression Recognition Trained Using Noiseless Data in 35 Training Poses (tp) from the **BU3DFE** Data Set, and Tested in a Subject-Independent Manner Using Data in 247 Test Poses (tp and non-tp (ntp)) from the **BU3DFE** Dataset, and Corrupted by Different Levels of Noise UNIF $\sim [-\sigma, \sigma]$, with $\sigma = 0, 2, 4$ Pixels (Where 10 Percent of Interocular Distance for the Average Registered Frontal-Pose Face in the **BU3DFE** Dataset Is Approximately 5 Pixels)

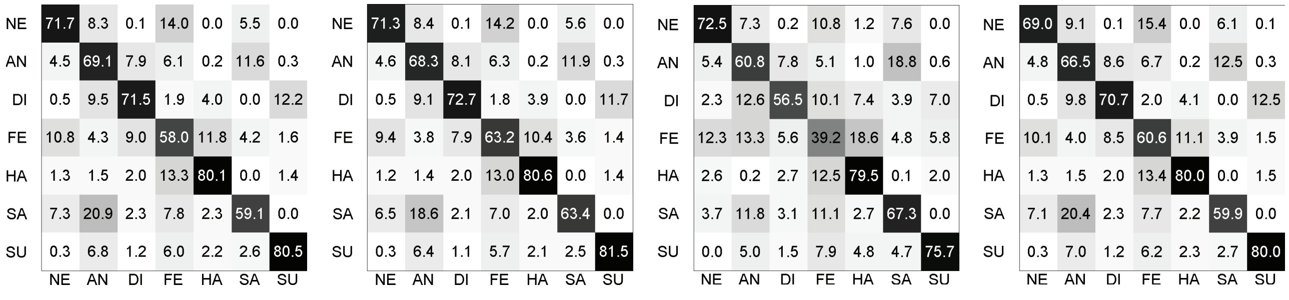
Method	RR (%) ($\sigma = 0$)		RMSE (in pixel) ($\sigma = 0$)		RR (%) ($\sigma = 2$)		RMSE (in pixel) ($\sigma = 2$)		RR (%) ($\sigma = 4$)		RMSE (in pixel) ($\sigma = 4$)	
	tp	ntp	tp	ntp	tp	ntp	tp	ntp	tp	ntp	tp	ntp
Pose-wise												
PW-SVM *(balanced)*	70.5	68.2	-	-	68.9	67.3	-	-	66.7	65.0	-	-
PDM & F-SVM *(balanced)*												
2D-PDM	59.5	59.1	3.18	3.21	57.2	56.7	3.33	3.45	55.2	54.1	3.58	3.62
3D-PDM	62.1	62.3	2.70	2.67	61.8	61.0	2.78	2.89	59.9	59.2	3.12	3.09
Regression *(balanced)* & F-SVM												
LR - DB	64.3	63.1	2.08	2.12	60.1	59.9	2.45	2.52	56.1	54.3	2.85	2.91
SVR - DB	68.7	68.1	1.60	1.63	67.9	67.0	1.92	2.06	66.7	65.2	2.19	2.21
TWINGPR - DB	70.1	68.5	1.18	1.29	68.2	67.8	1.63	1.79	66.4	66.5	2.12	2.27
MTGPR - DB	67.8	66.5	1.36	1.45	66.7	66.1	1.50	1.61	65.2	64.4	2.09	2.13
MTGPR - VB	68.1	67.2	1.32	1.40	67.1	66.8	1.48	1.58	65.4	65.1	1.98	2.01
IGPR - DB	71.9	70.5	1.25	1.29	69.5	68.4	1.50	1.61	67.4	66.9	1.85	1.88
IGPR - VB	72.0	69.4	1.22	1.26	68.3	67.6	1.51	1.75	67.1	66.0	1.87	1.95
SGPR - DB	71.6	70.1	1.30	1.34	69.9	68.8	1.53	1.69	69.0	68.9	1.71	1.82
SGPR - VB	71.8	69.8	1.19	1.26	69.0	68.8	1.59	1.76	68.3	67.2	1.78	1.89
CIGPR	**72.9**	**72.2**	**1.01**	1.15	70.2	69.0	**1.34**	**1.42**	68.1	67.7	1.72	1.80
CSGPR	72.6	71.5	1.05	**1.11**	**70.5**	**69.4**	1.37	1.45	**69.9**	**69.7**	**1.64**	**1.71**
Regression *(imbalanced)* & F-SVM												
LR - DB	57.6	56.1	2.28	2.43	54.1	53.1	2.79	2.81	52.7	52.2	3.01	3.11
SVR - DB	60.3	60.1	1.85	1.87	59.0	58.2	2.03	2.17	57.1	57.0	2.43	2.55
TWINGPR - DB	63.7	62.5	1.45	1.60	59.0	58.8	2.01	2.11	58.6	57.9	2.33	2.42
MTGPR - DB	63.1	62.6	1.47	1.58	61.7	61.1	2.07	2.19	60.1	59.5	2.73	2.81
MTGPR - VB	63.4	62.9	1.41	1.53	61.9	61.4	1.98	2.18	60.6	60.1	2.67	2.71
IGPR - DB	65.1	64.6	1.52	1.61	62.5	62.0	2.01	2.10	60.0	59.8	2.52	2.58
IGPR - VB	64.9	64.3	1.41	1.57	62.1	61.8	2.01	2.11	60.2	59.2	2.58	2.60
SGPR - DB	64.4	63.9	1.59	1.66	62.7	61.9	1.98	2.07	60.8	60.0	2.40	2.49
SGPR - VB	64.5	64.4	1.52	1.63	63.1	62.1	1.97	2.04	61.2	60.9	2.44	2.50
CIGPR	**71.5**	**70.2**	**1.09**	**1.22**	69.8	67.9	1.51	1.72	68.4	67.7	1.97	2.01
CSGPR	71.1	69.2	1.15	1.31	**70.0**	**68.2**	**1.43**	**1.68**	**69.1**	**68.9**	**1.86**	**1.92**

models), even when trained using a small number of training data. This is because they are able to efficiently explore training data from the neighboring poses, which cannot be accomplished by the other models. In what follows, we use $N = 200$ and $N = 500$ data-pairs to train the regression models and PW-SVMs, respectively, in order to keep them computationally tractable without a significant decrease in the models' performance.

So far, we evaluated the models using the noiseless data from the 35 training poses. We next test the robustness of the models to missing data and noisy data. To this end, we trained the regression models using balanced and imbalanced data (as explained below) sampled from the 35 training poses, and tested on noiseless and noise-corrupted data (with unknown head-pose) sampled from all 247 poses. The balanced dataset contained examples sampled (from four folds) per pose-pairs (nonfrontal poses and the frontal pose) and uniformly at random from all seven facial expressions. The imbalanced dataset was prepared as follows: Examples of Neutral facial expression, sampled (from four folds) at random, were used to train 50 percent of the pose-pairs, which were selected at random. For the rest of the pose-pairs, training examples were selected as in the balanced dataset. The fifth fold, containing examples of all facial expressions, was used to test such trained models, and this was repeated for all the folds. Furthermore, the test data were corrupted by adding noise to the locations of the facial landmarks, as explained in Table 3. For the 2D- and 3D-PDM, we selected 13 and 17 shape bases, respectively. The shape bases were chosen from the balanced dataset so that 95 percent of the energy was retained. Moreover, in the case of the 3D-PDM, we used the 3D facial points, and for the 2D-PDM we used the corresponding 2D facial points in the frontal pose. The PW-SVMs were trained using the balanced dataset, as in the previous experiment. In the case of "noncoupled" regression models, the predictions from different nonfrontal poses were combined using either DB or VB weighting, as described in Section 4.3. The latter approach was used only for MTGPR, IGPR, and SGPR since these models provide uncertainty in their predictions. To reduce the computational load of the coupled models, the parameters P_{min} (see Algorithm 1) and C_{min} (see Algorithm 2) were set to 0.1 and 0.8, respectively.[1] Also, the number of the coupled functions per pose was constrained to three.

Table 3 shows the comparative results. The performance of the 2D- and 3D-PDM is inferior to that of the PW-SVMs and the regression-based models. These results suggest that the employed face-shape-based models are unable to accurately recover facial motions caused by facial expressions in the presence of large head movements. This, in

[1] We used a small validation set containing examples of five randomly selected subjects to set P_{min} and C_{min}.



	NE	AN	DI	FE	HA	SA	SU
NE	71.7	8.3	0.1	14.0	0.0	5.5	0.0
AN	4.5	69.1	7.9	6.1	0.2	11.6	0.3
DI	0.5	9.5	71.5	1.9	4.0	0.0	12.2
FE	10.8	4.3	9.0	58.0	11.8	4.2	1.6
HA	1.3	1.5	2.0	13.3	80.1	0.0	1.4
SA	7.3	20.9	2.3	7.8	2.3	59.1	0.0
SU	0.3	6.8	1.2	6.0	2.2	2.6	80.5

(a) SGPR(bal.), RR=70.1%

	NE	AN	DI	FE	HA	SA	SU
NE	71.3	8.4	0.1	14.2	0.0	5.6	0.0
AN	4.6	68.3	8.1	6.3	0.2	11.9	0.3
DI	0.5	9.1	72.7	1.8	3.9	0.0	11.7
FE	9.4	3.8	7.9	63.2	10.4	3.6	1.4
HA	1.2	1.4	2.0	13.0	80.6	0.0	1.4
SA	6.5	18.6	2.1	7.0	2.0	63.4	0.0
SU	0.3	6.4	1.1	5.7	2.1	2.5	81.5

(b) CSGPR(bal.), RR=71.6%

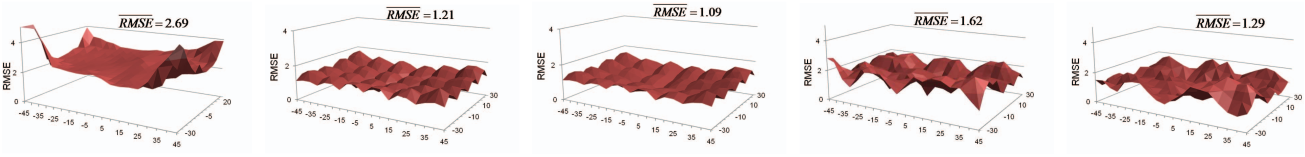
	NE	AN	DI	FE	HA	SA	SU
NE	72.5	7.3	0.2	10.8	1.2	7.6	0.0
AN	5.4	60.8	7.8	5.1	1.0	18.8	0.6
DI	2.3	12.6	56.5	10.1	7.4	3.9	7.0
FE	12.3	13.3	5.6	39.2	18.6	4.8	5.8
HA	2.6	0.2	2.7	12.5	79.5	0.1	2.0
SA	3.7	11.8	3.1	11.1	2.7	67.3	0.0
SU	0.0	5.0	1.5	7.9	4.8	4.7	75.7

(c) SGPR(imb.), RR=64.5%

	NE	AN	DI	FE	HA	SA	SU
NE	69.0	9.1	0.1	15.4	0.0	6.1	0.1
AN	4.8	66.5	8.6	6.7	0.2	12.5	0.3
DI	0.5	9.8	70.7	2.0	4.1	0.0	12.5
FE	10.1	4.0	8.5	60.6	11.1	3.9	1.5
HA	1.3	1.5	2.0	13.4	80.0	0.0	1.5
SA	7.1	20.4	2.3	7.7	2.2	59.9	0.0
SU	0.3	7.0	1.2	6.2	2.3	2.7	80.0

(d) CSGPR(imb.), RR=70.2%

Fig. 5. Confusion matrices for head-pose-invariant facial expression recognition obtained by (a) SGPR (balanced) and F-SVM, (b) CSGPR (balanced) and F-SVM, (c) SGPR (imbalanced) and F-SVM, and (d) CSGPR (imbalanced) and F-SVM. The methods were trained using noiseless data in 35 training poses from the **BU3DFE** dataset.



(a) 3D-PDM (b) SGPR (bal.) (c) CSGPR (bal.) (d) SGPR (imb.) (e) CSGPR (imb.)

Fig. 6. RMSE of the head-pose normalization in 247 tested poses attained by the 3D-PDM, and the SGPR and CSGPR models trained using the noiseless balanced/imbalanced data from the **BU3DFE** dataset.

turn, results in high RMSE and low RR attained by these two models. PW-SVM classifiers outperform the LR- and SVR-based methods, and perform similarly to the GPR-based methods when trained on the balanced data and tested on the noiseless data in discrete poses. However, they are less robust to noise and pose changes (i.e., test data from nontraining poses). Note that the results for the noiseless case and training poses differ from those shown in Table 2. This is caused by inaccuracies of the head pose estimation step. We also observe that TWINGPR is very sensitive to high levels of noise, which is reflected in its RMSE and RR. IGPR and SGPR show similar performance, with SGPR performing better in most cases in the recognition task. The performance of MTGPR in the target task is lower than that of IGPR. This is caused by the model overfitting due to the large number of the model parameters (due to the large number of the model outputs). On the other hand, the CIGPR- and CSGPR-based methods perform the best among the tested methods. Note also that their performance remains stable in the case of nontraining poses. It clearly suggests that these models are able to generalize well in the case of continuous change in head-pose despite the fact that they were trained on a limited set of training data in discrete poses. It also suggests that using only DB or VB weighting of the GPR-based models results in inferior performance compared to that attained by the proposed coupling scheme, which uses the CI fusion rule for combining the outputs of different mapping functions. We can also observe that when the test data are corrupted by the noise, there is an expected decline in performance of all the models. However, this is less pronounced in the CSGPR-based method than in the CIGPR-based method. This is a consequence of the base SGPR model being able to learn the *structure* in the model output, which is important in the presence of high levels of noise. Finally, in the case of the imbalanced dataset, the performance of the "non-coupled" models is substantially lower compared to that of the CIGPR and CSGPR models. This clearly shows the benefit of using the proposed coupling scheme. Since these two models exhibit similar performance, with CSGPR performing better in the case of noisy data and being computationally much less intense, in further experiments we evaluate CSGPR and use SGPR (VB) as the baseline model.

Fig. 5 shows the confusion matrices for facial expression recognition attained by the SGPR- and CSGPR-based methods. In contrast to the CSGPR-based method, the RRs of the SGPR-based method decrease considerably in the case of the imbalanced dataset compared to when this model is trained using the balanced dataset. However, the SGPR-based method outperforms the CSGPR-based method on the Neutral facial expression class (when trained using the imbalanced dataset). This is because, for some pose-pairs, the SGPR models are trained using data of Neutral facial expression only, and thus there is no need for their coupling. Still, the CSGPR-based method shows a better performance on average. Fig. 6 depicts changes in the RMSE of different models across the tested poses. As can be seen, the RMSE of the 3D-PDM increases rapidly in poses that are far from frontal, indicating that the used 3D-PDM model is unable to accurately recover the 3D face shape from the 2D points in

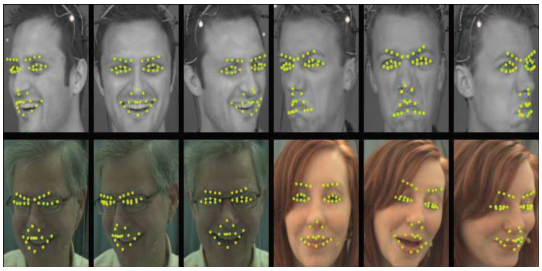


Fig. 7. Example images from the **MPFE** dataset (top) and the **Semaine** dataset (bottom) with the facial points automatically localized by the AAM [23].

TABLE 4
The Results of the State-of-the-Art Methods for Head-Pose Invariant Facial Expression Recognition on the **BU3DFE** Dataset

Method	Classifier	Features	Poses			Expressions		Recognition Rates			
			tilt	pan	total	number	levels	tp(bal.)	ntp(bal.)	tp(imbal.)	ntp(imbal.)
Hu et al. [49]	pose-wise svm	41 landmarks	-	(0°,+90°)	5	6	1, 2, 3, 4	66.7%	-	-	-
Moore and Bowden [9]	pose-wise svm	lgbp\lbp	-	(0°,+90°)	5	6	1, 2, 3, 4	71.1%	-	-	-
Hu et al. [27]	single knn	sift+lpp	-	(0°,+90°)	5	6	2, 3, 4	73.8%	-	-	-
Zheng et al. [28]	pose-wise knn	83 landmarks+sift	-	(0°,+90°)	5	6	1, 2, 3, 4	78.5%	-	-	-
Zheng et al. [10]	single linear	sift + bda\gmm	(-30°,+30°)	(-45°,+45°)	35	6	4	68.3%	-	-	-
Tang et al. [11]	pose-wise svm	sift + hmm	(-30°,+30°)	(-45°,+45°)	35	6	4	75.3%	-	-	-
SGPR (lev. 3)	frontal svm	39 landmarks	(-30°,+30°)	(-45°,+45°)	247	7	3	68.2%	65.4%	61.5%	61.4%
CSGPR (lev. 3)	frontal svm	39 landmarks	(-30°,+30°)	(-45°,+45°)	247	7	3	68.7%	67.1%	68.0%	65.3%
SGPR (lev. 4)	frontal svm	39 landmarks	(-30°,+30°)	(-45°,+45°)	247	7	4	75.4%	74.2%	67.5%	67.4%
CSGPR (lev. 4)	frontal svm	39 landmarks	(-30°,+30°)	(-45°,+45°)	247	7	4	76.5%	76.1%	74.2%	72.9%

these poses. The 3D-PDM and the LR-based method show similar performance on average. Also, their performance is inferior to that obtained by the CSGPR-based method, which generalizes well even in the nontraining poses.

Table 4 gives an overview of the results obtained on the BU3DFE dataset by the proposed CSGPR-based method and previously proposed methods for head-pose-invariant facial expression recognition. When studying the results shown in Table 4, the following should be considered. First, the methods proposed in [49], [9], [27], and [28] were trained/tested on a small set of discrete poses containing only pan rotations. In other words, they do not deal with large head-pose changes. Second, the methods proposed in [49], [9], [27], [28], and [11] are person specific since they use the neutral frame in the feature preprocessing step. Therefore, they are inapplicable to real-world scenarios. The method proposed in this paper and the methods proposed in [28] and [11] are the only ones that consider the "full" range of poses, including pan and tilt rotations with a significant part of the face remaining visible. Yet, the methods in [28] and [11] were evaluated on a set of discrete poses used for training, so it is not clear how these methods would perform in nontraining poses. On the other hand, the proposed CSGPR method and the baseline SGPR method (C/SGPR methods) were evaluated on both training and nontraining poses, and using balanced and imbalanced datasets. Furthermore, most of the methods in Table 4 were trained pose-wise, and hence could not deal with missing facial expressions (i.e., the imbalanced data), as opposed to the C/SGPR-based methods. For the C/SGPR-based methods, in Table 4, we report the results per expression levels 3 and 4 separately so that they can be compared with the results of the other methods, which usually consider only expression level 4. Note that Table 3 (the noiseless case) shows the average of the results for the both levels.

5.3 Experiments on Real-Image Data

In the experiments on the real-image data from the MultiPIE dataset, we prepared the imbalanced datasets as follows: For pose $(0^\circ, -30^\circ)$ and for facial expression of, e.g., Surprise, we removed all examples of this facial expression from the pose in question, and kept the examples of all four facial expressions in the two remaining (nonfrontal) poses. This was repeated for each facial expression and nonfrontal pose. Such datasets were then used to train the SGPR models for each pair of a nonfrontal and the frontal pose, which, in the case of the CSGPR model, were then coupled.

Table 5 shows the performance of the C/SGPR-based methods trained using the balanced and imbalanced data from the MulitPIE dataset. In the former case, the testing was done at once, i.e., by using the examples of all facial expressions in all nonfrontal poses. The methods trained on the imbalanced datasets were tested using only the examples of the missing facial expression in the target pose. As can be seen from Table 5, both methods perform similarly when the balanced datasets are used. This is especially the case for facial expressions of Neutral and Disgust. We attribute this to the fact that, in the case of the perfectly balanced dataset, some of the coupled functions in the CSGPR model add noise to the final prediction in the frontal pose as a consequence of the registration process. In the case of the imbalanced dataset, the CSGPR-based method outperforms the SGPR-based method. Again, this is due to the SGPR-based method being unable to generalize well beyond the training data per training pairs of poses.

We further compare the performance of the C/SGPR-based methods using the MPFE dataset. We include here the results attained by the AAM-based method for pose-invariant facial expression recognition. For this, we used the AAM [23], i.e., its Candide model (being the 3D Active Shape Model part of the AAM), to perform the head-pose normalization. The latter was attained by rotating the Candide model to the frontal pose, where the 2D (pose-normalized) facial points were obtained from the corresponding 3D points. The manual initialization of the Candide model in the frontal pose and the corresponding 2D points obtained from the initialization step were used as the ground truth when computing the RMSE and to train the F-SVM. Table 6 summarizes the average results per expression. As can be seen, the CSGPR-based method outperforms the AAM (Candide) in the task of head-pose

TABLE 5
RMSE and RR (per Expression) Attained by the SGPR- and CSGPR-Based Methods, Trained/Tested Using Balanced (bal.) and Imbalanced (imb.) Data from the **MulitPIE** Dataset

	RR (%)				RMSE (in pixel)			
	SGPR (bal.)	CSGPR (bal.)	SGPR (imb.)	CSGPR (imb.)	SGPR (bal.)	CSGPR (bal.)	SGPR (imb.)	CSGPR (imb.)
NE	93.7	93.8	84.4	89.5	1.45	1.39	2.35	1.86
DI	92.0	91.7	75.7	82.1	1.60	1.52	2.91	1.91
JO	93.9	95.6	84.2	90.1	1.59	1.51	2.85	1.88
SU	96.6	98.1	82.8	88.7	1.65	1.59	2.95	2.31
Av.	94.1	**94.8**	81.8	**87.6**	1.57	**1.50**	2.76	**1.99**

TABLE 6
RMSE and RR (per Expression) Attained by the AAM (Candide) and the SGPR- and CSGPR-Based Methods, Trained Using Balanced (bal.) and Imbalanced (imb.) Data from the **MPFE** Dataset

	RR (%)					RMSE (in pixel)				
	AAM (Cand.)	SGPR (bal.)	CSGPR (bal.)	SGPR (imb.)	CSGPR (imb.)	AAM (Cand.)	SGPR (bal.)	CSGPR (bal.)	SGPR (imb.)	CSGPR (imb.)
NE	72.2	83.4	85.0	73.2	79.1	3.51	1.85	1.61	2.85	2.55
AN	54.1	72.5	73.2	59.6	64.4	3.24	1.98	1.82	3.11	2.95
DI	58.7	73.0	74.8	62.7	70.1	4.13	2.25	2.11	3.80	3.62
FE	60.4	68.3	69.9	59.5	64.6	3.44	2.20	2.30	3.15	2.83
JO	72.9	87.2	89.1	77.0	83.2	4.21	2.38	2.42	3.45	3.15
SA	57.2	68.9	70.2	60.1	63.7	3.65	2.01	1.90	3.60	3.09
SU	78.1	88.5	91.5	76.7	85.2	4.42	2.61	2.51	3.45	3.11
Av.	64.8	77.4	**79.1**	67.0	**73.0**	3.8	2.18	**2.09**	3.34	**3.04**

normalization. This is because the pose normalization based on the Candide model is more susceptible to tracking errors since, in contrast to the CSGPR-based method, no training data are used to smooth out the noise in its output. Also, the rotation matrix used to bring the Candide model to the frontal pose is learned based on the pose-estimation provided by the AAM [23]. So, the inaccuracy of the pose estimation also degrades the performance of this model. In the case of the imbalanced dataset, the CSGPR-based method largely outperforms the AAM- and SGPR-based methods. However, there is a decline in performance attained by all the methods. In the case of C/SGPR this is expected since they are trained using not only the imbalanced data but also the data of only two subjects (a threefold person-independent cross-validation procedure was applied in this experiment).

We also evaluated the proposed method on spontaneously displayed facial expressions from the Semaine data set [46]. Specifically, we performed cross-database evaluation where the C/SGPR-based methods were trained using the MultiPIE and the MPFE datasets, and tested using the Semaine dataset. Table 7 shows that the C/SGPR-based methods generalize well, with CSGPR outperforming the base SGPR in all the tasks, despite the fact that they were trained using a different dataset from the one used for testing. Note also that the C/SGPR-based methods trained on the MPFE dataset outperform those trained on the MultiPIE dataset. This is due to the difference in facial point localization, which, in the case of the MultiPIE dataset, was done manually and, in the case of the MPFE and Semaine datasets, was done automatically. Consequently, the C/SGPR-based methods trained on the MultiPIE dataset were more sensitive to noise in test data.

6 Conclusion

We have proposed a method for head-pose-invariant facial expression recognition that is based on 2D geometric features. We have shown that the proposed CSGPR model for head-pose normalization outperforms the state-of-the-art regression-based approaches to head-pose normalization, the 2D- and 3D-PDMs and the online AAM. In contrast to the existing pose-invariant facial expression recognition methods, the proposed method can deal with missing data (i.e., facial expression categories that were not available in certain nonfrontal poses during training).

TABLE 7
RR for Facial Expressions of Laughter and Speech Attained by the AAM (Candide), and the SGPR- and CSGPR-Based Methods, Trained Using Balanced (bal.) and Imbalanced (imb.) Data from the **MPFE** and **MultiPIE** Data Sets, and Tested on the **Semaine** Dataset

	RR (%)				
	AAM (Semaine)	SGPR (MPFE)	CSGPR (MPFE)	SGPR (MultiPIE)	CSGPR (MultiPIE)
Laughter	64.1	80.4	83.2	69.5	77.1
Speech	93.2	89.8	94.8	85.7	90.3
Av.	78.6	85.1	**89.0**	77.6	**83.7**

Acknowledgments

This work is funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Ioannis Patras is supported in part by the Engineering and Physical Sciences Research Council Grant EP/G033935/1.

References

[1] M. Pantic, A. Nijholt, A. Pentland, and T. Huang, “Human-Centred Intelligent Human-Computer Interaction (HCI2): How Far Are We from Attaining It?” *Int'l J. Autonomous and Adaptive Comm. Systems,* vol. 1, no. 2, pp. 168-187, 2008.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, “Social Signal Processing: Survey of an Emerging Domain,” *Image and Vision Computing J.,* vol. 27, no. 12, pp. 1743-1759, 2009.

[3] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 1, pp. 39-58, Jan. 2009.

[4] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I.R. Fasel, and J.R. Movellan, “Automatic Recognition of Facial Actions in Spontaneous Expressions,” *J. Multimedia,* vol. 1, no. 6, pp. 22-35, 2006.

[5] Y. Zhang and Q. Ji, “Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 699-714, May 2005.

[6] Y. Tong, W. Liao, and Q. Ji, “Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 10, pp. 1683-1699, Oct. 2007.

[7] Z. Zhu and Q. Ji, “Robust Real-Time Face Pose and Facial Expression Recovery,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 681-688, 2006,

[8] Y. Tong, J. Chen, and Q. Ji, “A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 2, pp. 258-273, Feb. 2010.

[9] S. Moore and R. Bowden, “Local Binary Patterns for Multi-View Facial Expression Recognition,” *Computer Vision and Image Understanding,* vol. 115, no. 4, pp. 541-558, 2011.

[10] W. Zheng, H. Tang, Z. Lin, and T.S. Huang, “Emotion Recognition from Arbitrary View Facial Images,” *Proc. European Conf. Computer Vision,* pp. 490-503, 2010,

[11] H. Tang, M. Hasegawa-Johnson, and T.S. Huang, “Non-Frontal View Facial Expression Recognition Based on Ergodic Hidden Markov Model Supervectors,” *Proc. Int'l Conf. Multimedia and Expo,* pp. 1202-1207, 2010,

[12] M.J. Black and Y. Yacoob, “Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion,” *Int'l J. Computer Vision,* vol. 25, pp. 23-48, 1997.

[13] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, “Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates,” *Int'l J. Computer Vision,* vol. 83, no. 2, pp. 178-194, 2009.

[14] P. Ekman and W.V. Friesen, *Unmasking the Face: A Guideline to Recognising Emotions from Facial Clues,* vol. 3. Prentice Hall, 1978.

[15] C.M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2007.

[16] Y. Sun and L. Yin, "Facial Expression Recognition Based on 3D Dynamic Range Model Sequences," *Proc. European Conf. Computer Vision,* pp. 58-71, 2008.

[17] L.A. Jeni, A. Lrincz, T. Nagy, Z. Palotai, J. Sebk, Z. Szab, and D. Takcs, "3D Shape Estimation in Video Sequences Provides High Precision Evaluation of Facial Expressions," *Image and Vision Computing,* vol. 30, pp. 785-795, 2012.

[18] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D Facial Expression Recognition Based on Primitive Surface Feature Distribution," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1399-1406, 2006,

[19] R.E. Kaliouby and P. Robinson, "Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops,* p. 154. 2004,

[20] W.-K. Liao and I. Cohen, "Belief Propagation Driven Method for Facial Gestures Recognition in Presence of Occlusions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops,* p. 158. 2006,

[21] S. Lucey, A.B. Ashraf, and J. Cohn, *Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face.* I-Tech Education & Publishing, 2007.

[22] J. Sung and D. Kim, "Real-Time Facial Expression Recognition Using STAAM and Layered GDA Classifier," *Image and Vision Computing J.,* vol. 27, no. 9, pp. 1313-1325, 2009.

[23] F. Dornaika and J. Orozco, "Real Time 3D Face and Facial Feature Tracking," *J. Real-Time Image Processing,* vol. 2, no. 1, pp. 35-44, 2007.

[24] R. Gross, I. Matthews, and S. Baker, "Generic vs. Person Specific Active Appearance Models," *Image and Vision Computing J.,* vol. 23, pp. 1080-1093, 2005.

[25] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally Linear Regression for Pose-Invariant Face Recognition," *IEEE Trans. Image Processing,* vol. 16, no. 7, pp. 1716-1725, July 2007.

[26] M.A.O. Vasilescu and D. Terzopoulos, "Multilinear Analysis of Image Ensembles: Tensorfaces," *Proc. European Conf. Computer Vision,* pp. 447-460, 2002.

[27] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T.S. Huang, "Multi-View Facial Expression Recognition," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 1-6, 2008,

[28] W. Zheng, H. Tang, Z. Lin, and T. Huang, "A Novel Approach to Expression Recognition from Non-Frontal Face Images," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1901-1908, 2009.

[29] O. Rudovic, I. Patras, and M. Pantic, "Coupled Gaussian Process Regression for Pose-Invariant Facial Expression Recognition," *Proc. European Conf. Computer Vision,* pp. 350-363, 2010.

[30] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial Point Detection Using Boosted Regression and Graph Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2729-2736, 2010.

[31] E. Murphy-Chutorian and M.M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 4, pp. 607-626, Apr. 2009.

[32] M.F. Valstar and M. Pantic, "Fully Automatic Recognition of the Temporal Phases of Facial Actions," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 42, no. 1, pp. 28-43, Feb. 2012.

[33] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett, "New Support Vector Algorithms," *Neural Computation,* vol. 12, pp. 1207-1245, 2000.

[34] K. Grochow, S.L. Martin, A. Hertzmann, and Z. Popović, "Style-Based Inverse Kinematics," *Proc. ACM Int'l Conf. Computer Graphics and Interactive Techniques,* pp. 522-531, 2004.

[35] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning.* The MIT Press, 2005.

[36] S.J. Julier and J.K. Uhlmann, "A Non-Divergent Estimation Algorithm in the Presence of Unknown Correlations," *Proc. Am. Control Conf.,* pp. 2369-2373, 1997.

[37] V. Tresp and M. Taniguchi, "Combining Estimators Using Non-Constant Weighting Functions," *Proc. Neural Information Processing Systems Conf.,* pp. 419-426, 1995.

[38] V. Tresp, "A Bayesian Committee Machine," *Neural Computing,* vol. 12, no. 11, pp. 2719-2741, 2000.

[39] P. Boyle and M. Frean, "Dependent Gaussian Processes," *Neural Information Processing Systems,* pp. 217-224, 2005.

[40] K. Yu, V. Tresp, and A. Schwaighofer, "Learning Gaussian Processes from Multiple Tasks," *Proc. 22nd Int'l Conf. Machine Learning,* pp. 1012-1019, 2005.

[41] E.V. Bonilla, K.M.A. Chai, and C.K.I. Williams, "Multi-Task Gaussian Process Prediction," *Neural Information Processing Systems,* 2008.

[42] M. Alvarez and N. Lawrence, "Sparse Convolved Multiple Output Gaussian Processes," *Proc. Neural Information Processing Systems Conf.,* pp. 57-64, 2008.

[43] L. Bo and C. Sminchisescu, "Twin Gaussian Processes for Structured Prediction," *In'l J. Computer Vision,* vol. 87, nos. 1/2, pp. 28-52, 2010.

[44] S. Yu, B. Krishnapuram, R. Rosales, and R.B. Rao, "Bayesian Co-Training," *J. Machine Learning Research,* vol. 12, pp. 2649-2680, 2011.

[45] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-Pie," *Image and Vision Computing J.,* vol. 28, no. 5, pp. 807-813, 2010.

[46] G. McKeown, M.F. Valstar, R. Cowie, and M. Pantic, "The Semaine Corpus of Emotionally Coloured Character Interactions," *Proc. Int'l Conf. Multimedia and Expo,* pp. 1079-1084, 2010.

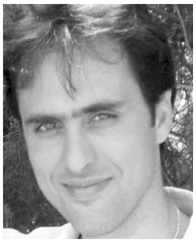
[47] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intelligent Systems and Technology,* vol. 2, pp. 1-27, 2011.

[48] T. Cootes and C. Taylor, "Active Shape Models—Smart Snakes," *Proc. British Machine Vision Conf.,* pp. 266-275, 1992.

[49] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T. Huang, "A Study of Non-Frontal-View Facial Expressions Recognition," *Proc. Int'l Conf. Pattern Recognition,* pp. 1-4, 2008.

Ognjen Rudovic received the BSc degree in automatic control from the Faculty of Electrical Engineering, University of Belgrade, Serbia, in 2007, and the MSc degree in computer vision from the Computer Vision Center (CVC), Universitat Autonoma de Barcelona, Spain, in 2008. Currently, he is working toward the PhD degree in the Department of Computing at Imperial College London, United Kingdom. His research interests include automatic affect recognition, machine learning, and computer vision. He is a student member of the IEEE.

Maja Pantic is a professor in affective and behavioral computing in the Department of Computing at Imperial College London, United Kingdom, and in the Department of Computer Science at the University of Twente, The Netherlands. She currently serves as the editor in chief of *Image and Vision Computing Journal* and as an associate editor for both the *IEEE Transactions on Systems, Man, and Cybernetics Part B* and the *IEEE Transactions on Pattern Analysis and Machine Intelligence*. She has received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She is a fellow of the IEEE.



Ioannis (Yiannis) Patras received the BSc and MSc degrees in computer science from the Computer Science Department, University of Crete, Heraklion, Greece, in 1994 and 1997, respectively, and the PhD degree from the Department of Electrical Engineering, Delft University of Technology, The Netherlands, in 2001. He is a senior lecturer in the School of Electronic Engineering and Computer Science, Queen Mary University London, United Kingdom. His current research interests include computer vision and pattern recognition, with emphasis on the analysis of human motion, including the detection, tracking, and understanding of facial and body gestures and their applications in multimedia data management, multimodal human computer interaction, and visual communication. He is an associate editor of the *Image and Vision Computing Journal*. He is a senior member of the IEEE.