Shared Gaussian Process Latent Variable Model for Multi-view Facial Expression Recognition

Stefanos Eleftheriadis[1], Ognjen Rudovic[1], and Maja Pantic[1,2]

[1] Comp. Dept., Imperial College London, UK
[2] EEMCS, University of Twente, The Netherlands

Abstract. Facial-expression data often appear in multiple *views* either due to head-movements or the camera position. Existing methods for multi-view facial expression recognition perform classification of the target expressions either by using classifiers learned *separately* for each view or by using a single classifier learned for all views. However, these approaches do not explore the fact that multi-view facial expression data are different manifestations of the same facial-expression-related latent content. To this end, we propose a Shared Gaussian Process Latent Variable Model (SGPLVM) for classification of multi-view facial expression data. In this model, we first learn a discriminative manifold shared by multiple views of facial expressions, and then apply a (single) facial expression classifier, based on k-Nearest-Neighbours (kNN), to the shared manifold. In our experiments on the MultiPIE database, containing real images of facial expressions in multiple views, we show that the proposed model outperforms the state-of-the-art models for multi-view facial expression recognition.

1 Introduction

Facial expression recognition has attracted significant research attention because of its usefulness in many applications, such as human-computer interaction, security and analysis of social interactions [1,2]. Most existing methods deal with imagery in which the depicted persons are relatively still and exhibit posed expressions in a nearly frontal pose [3]. However, most real-world applications relate to spontaneous interactions (*e.g.*, meeting summarization, political debates analysis, etc.), in which the assumption of having immovable subjects is unrealistic. This calls for a joint analysis of facial expressions and head-poses. 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.

To date, only a few works that deal with multi-view facial expression data have been proposed. These methods can be divided into three groups depending on how they deal with the variation in head pose of the subjects depicted in the images. In what follows, we review the existing models. The first group of the methods perform *pose-wise* facial expression recognition. In [4], the authors used Local Binary Patterns (LBP) [5] (and its variants) to perform a two-step facial expression classification. In the first step, they select the closest head-pose to the (discrete) training pose by using the SVM classifier. Once the pose is known, they apply the pose-specific SVM to perform facial-expression classification in the selected pose. In [6], different appearance features (SIFT, HoG,

G. Bebis et al. (Eds.): ISVC 2013, Part I, LNCS 8033, pp. 527–538, 2013.

LBP) are extracted around the locations of characteristic facial points, and used to train various pose-specific classifiers. Similarly, [7] used pose-wise 2D AAMs [8] to locate a set of characteristic facial points, which then were used as the input features for the classifiers in each pose. Another group of approaches ([9,10]) first perform head-pose normalization, and then apply facial expression classification in the canonical pose, usually chosen to be the frontal. The main downside of all these approaches is that they ignore correlations across different poses, which makes them suboptimal for the target task. Furthermore, by learning separate classifiers for each view, the *pose-wise* methods may give inconsistent classification of facial expressions across the views. As shown by [4,6], recognition of some facial expressions can be performed better in certain poses. Hence, finding a joint feature space for multi-view facial expression recognition may improve overall performance of the model. This is in part explored in [11,12], where the authors learn a single classifier for data from multiple poses. Specifically, [11] use variants of dense SIFT [13] features extracted from multi-view facial expression images, an attempt to align the data from different poses during the feature extraction step. Likewise, [12] used the Generic Sparse Coding scheme ([14]) to learn a dictionary that sparsely encodes the SIFT features extracted from facial images in different views. However, high variation in facial features extracted from different views increases the complexity of the learned classifier significantly since it attempts to simultaneously deal with variation in head-pose and facial expressions.

Note that none of the works mentioned above explores the fact that the multi-view data are usually different manifestations of the same (latent) facial-expression-specific content. To this end, in this paper we propose a discriminative Shared Gaussian Process Latent Variable Model (DS-GPLVM) for multi-view facial expression recognition. In the proposed model, we learn a joint low-dimensional facial-expression manifold of the expression data from multiple views. To attain good classification of the target facial expressions in the shared space, we place a discriminative prior informed by the expression labels over the manifold. This model is based on the discriminative GPLVM (D-GPLVM) [15], proposed for non-linear dimensionality reduction and classification of data from a single observation space. We generalize this model so that it can simultaneously handle multiple observation spaces. Although the proposed model is applicable to a variety of learning tasks (multi-view classification, multiple-feature fusion, etc.), in this paper we limit our consideration to multi-view facial expression recognition. The outline of the proposed model is given in Fig. 1.

The remainder of the paper is organized as follows. We give a short overview of the base GPLVM and the D-GPLVM in Sec. 2. In Sec. 3, we introduce the proposed Discriminative Shared Gaussian Process Latent Variable Model for multi-view facial expression recognition. Sec. 4 shows the results of the experiments conducted, and Sec. 5 concludes the paper.

2 Gaussian Process Latent Variable Models (GPLVM)

In this section, we give a brief overview of the GPLVM [16], commonly used for learning complex low-dimensional data manifolds. We then introduce two types of discriminative priors for the data-manifold, which are used to obtain the discriminative GPLVM.

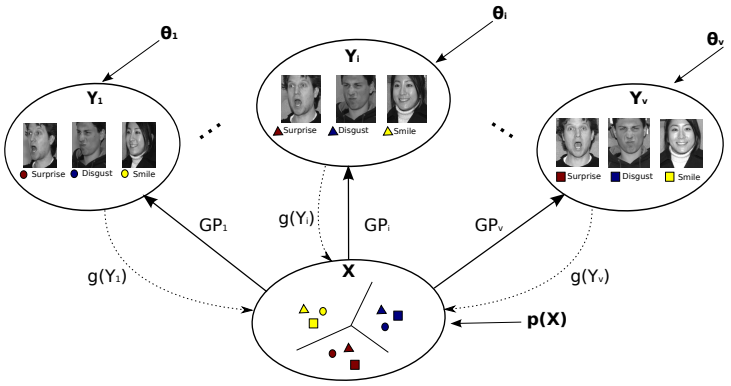


Fig. 1. The overview of the proposed DS-GPLVM. The discriminative shared manifold $\mathbf{X}$ of facial images from different views ($\mathbf{Y}_i$, $i = 1 \dots V$) is learned using the framework of shared GPs (GP$_i$). The class separation in the shared manifold is enforced by the discriminative shared prior $p(\mathbf{X})$, informed by the data labels. During inference, the facial images from different views are projected onto the shared manifold by using the kernel-based regression, learned for each view separately ($g(\mathbf{Y}_i)$). The classification of the query image is performed using the k-NN.

2.1 Gaussian Process Latent Variable Model (GPLVM)

The GPLVM [16] is a probabilistic model for non-linear dimensionality reduction. It learns a low dimensional latent space $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T \in \mathcal{R}^{N \times q}$, with $q \ll D$, corresponding to the high dimensional observation space $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^T \in \mathcal{R}^{N \times D}$. The mapping between the latent and observation space is modeled using the framework of Gaussian Processes (GP) [17]. Specifically, by using the covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ of the GP, the likelihood of the observed data, given the latent coordinates, is

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \frac{1}{\sqrt{(2\pi)^{ND}|\mathbf{K}|^D}} \exp(-\frac{1}{2}\text{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T)), \tag{1}$$

where $\mathbf{K}$ is the kernel matrix with the elements given by the covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$. The covariance function is usually chosen as the sum of the Radial Basis Function (RBF) kernel, and the bias and noise terms

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp(-\frac{\theta_2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2) + \theta_3 + \frac{\delta_{i,j}}{\theta_4}, \tag{2}$$

where $\delta_{i,j}$ is the Kronecker delta function and $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ are the kernel parameters [17]. The latent space X is obtained by using the mean of the posterior distribution

$$p(\mathbf{X}, \theta|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \theta)p(\mathbf{X}) \tag{3}$$

where the flat Gaussian prior is imposed on the latent space to prevent the GPLVM from placing the latent points infinitely apart. The learning of the latent space is accomplished by minimizing the negative log likelihood of the posterior in (3), w.r.t. $\mathbf{X}$, and it is given by

$$L = \frac{D}{2}\ln|\mathbf{K}| + \frac{1}{2}\text{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T) - \log(P(\mathbf{X})). \tag{4}$$

2.2 Discriminative Gaussian Process Latent Variable Model (D-GPLVM)

Note that the GPLVM is an unsupervised method for dimensionality reduction, and, as such, it is not optimal for the classification tasks. However, due to its probabilistic formulation, this model can easily be adapted for classification by placing a discriminative prior over the latent space, instead of the flat Gaussian prior. This has been firstly explored in [15], where a prior based on Linear Discriminant Analysis (LDA) is used. LDA tries to maximize between-class separability and minimize within-class variability by maximizing

$$J(\mathbf{X}) = \text{tr}(\mathbf{S}_w^{-1}\mathbf{S}_b), \tag{5}$$

where $\mathbf{S}_w$ and $\mathbf{S}_b$ are the within- and between-class matrices:

$$\mathbf{S}_w = \sum_{i=1}^{L} \frac{N_i}{N} \left[\frac{1}{N_i} \sum_{k=1}^{N_i} (x_k^{(i)} - \mathbf{M}_i)(x_k^{(i)} - \mathbf{M}_i)^T \right], \tag{6}$$

$$\mathbf{S}_b = \sum_{i=1}^{L} \frac{N_i}{N} (\mathbf{M}_i - \mathbf{M}_0)(\mathbf{M}_i - \mathbf{M}_0)^T \tag{7}$$

where $\mathbf{X}^{(i)} = [x_1^{(i)}, \ldots, x_{N_i}^{(i)}]$ are the N_i training points from class i, $\mathbf{M}_i$ is the mean of the elements of class i, and $\mathbf{M}_0$ is the mean of the training points from all the classes. The function in (5) is then used to define a prior over the latent positions, which is given by

$$p(\mathbf{X}) = \frac{1}{Z_d} \exp\left\{ -\frac{1}{\sigma_d^2} J^{-1} \right\}, \tag{8}$$

where Z_d is a normalization constant and σ_d represents a global scaling of the prior. By replacing the Gaussian prior in (3) with the prior in (8) we obtain the Discriminative GPLVM [15]. The authors also proposed a non-linear version of the prior based on Generalized Discriminant Analysis (GDA). Note, however, that in both cases, the dimension of the latent space is at most C, where C is the number of classes.

The limitation of the above-defined discriminative prior is overcome in the GP Latent Random Field (GPLRF) model [18], where the authors proposed a prior based on Gaussian Markov Random Field (GMRF) [19]. Specifically, an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed, where $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$ is the node set, with node V_i corresponding to a training example $\mathbf{x}_i$. $\mathcal{E} = \{V_i, V_j\}_{i,j=1\ldots N}$ is the edge set with $\mathbf{x}_i$ and $\mathbf{x}_j$ belonging to the same class, and $i \neq j$. By pairing each node with the random vector $\mathbf{X}_{*k} = (\mathbf{X}_{1k}, \mathbf{X}_{2k}, \ldots, \mathbf{X}_{Nk})^T$ (for $k = 1, 2, \ldots, q$), we obtain a Markov random field over the latent space. We next associate each edge with a weight 1 to build a weight matrix

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j,\ i \neq j, \text{ belong to the same class} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The graph Laplacian matrix is then defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$. Finally, using $\mathbf{L}$, the discriminative GMRF prior is defined as

$$p(\mathbf{X}) = \prod_{k=1}^{q} p(\mathbf{X}_{*k}) = \frac{1}{Z_q} \exp\left[-\frac{\beta}{2} \text{tr}(\mathbf{X}^T \mathbf{L} \mathbf{X}) \right], \tag{10}$$

where Z_q is a normalization constant and $\beta > 0$ is a scaling parameter. The term $tr(\mathbf{X}^T\mathbf{L}\mathbf{X})$ in the discriminative prior in (10) reflects the sum of the distances between the latent points from the same class, resulting in the latent points from the same class having higher $p(\mathbf{X})$. Thus, with the GMRF prior in (3) we penalize more the latent spaces that are less discriminative in terms of the target classes.

3 Discriminative Shared GPLVM (DS-GPLVM)

The D-GPLVM from the previous section is applicable only to a single observation space. In this section, we generalize the D-GPLVM so that it can simultaneously learn a discriminative manifold of multiple observation spaces. This is attained by using the framework for learning the shared manifold of multiple observation spaces ([20,21]), and by introducing a multi-view discriminative prior for the shared manifold. We assume in our approach that the multiple observation spaces are dependent, so that a single discriminative shared manifold can be used for their reconstruction. In the case of multi-view facial expression data, we expect this assumption to hold, since the appearance of facial expressions captured at different views changes mainly because of the view variation. Thus, the goal of learning their shared manifold is to perform the simultaneous alignment of facial-expression-related features from different views.

3.1 Shared-GPLVM

Recently, the Shared-GPLVM [20,21,22] has been proposed for learning a shared latent representation $\mathbf{X}$ that captures the correlations among different sets of corresponding features $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_V\}$, where V is the number of different feature sets (in our case, different views). This is achieved by modifying the standard GPLVM so that it can learn V Gaussian Processes, each generating one observation space from the shared latent space. Specifically, the joint marginal likelihood of the set of the observation spaces is given by

$$p(\mathbf{Y}_1, \ldots, \mathbf{Y}_V|\mathbf{X}, \theta_s) = p(\mathbf{Y}_1|\mathbf{X}, \theta_{Y_1}) \ldots p(\mathbf{Y}_V|\mathbf{X}, \theta_{Y_V}), \tag{11}$$

where $\theta_s = \{\theta_{Y_1}, \ldots, \theta_{Y_V}\}$ are the kernel parameters for each observation space, and the kernel function is defined as in (2). The shared latent space $\mathbf{X}$ is then found by minimizing the joint negative log-likelihood penalized with the prior placed over the shared manifold, and is given by

$$L_s = \sum_v L^{(v)} - log(P(\mathbf{X})) \tag{12}$$

where $L^{(v)}$ is the negative log-likelihood of each of the observation spaces and is given by

$$L^{(v)} = \frac{D}{2} \ln |\mathbf{K}_v| + \frac{1}{2}\text{tr}(\mathbf{K}_v^{-1}\mathbf{Y}_v\mathbf{Y}_v^T) + \frac{ND}{2} \ln 2\pi, \tag{13}$$

where $\mathbf{K}_v$ is the kernel matrix associated with the input data $\mathbf{Y}_v$. As in GPLVM model, Shared-GPLVM uses the flat Gaussian prior for the latent positions.

3.2 Discriminative Shared-space Prior

To learn a discriminative shared space, we introduce a discriminative shared-space prior. Similarly as in the GMRF prior defined in (9) for the single view, we first construct the weight matrix $\mathbf{W}$ for each view but by using data-dependent weights, which are obtained by applying the heat kernel to the data from each view as

$$\mathbf{W}_{ij}^{(v)} = \begin{cases} \exp\left(-t^{(v)}\|\mathbf{y}_i^{(v)} - \mathbf{y}_j^{(v)}\|^2\right) & \text{if } \mathbf{y}_i \text{ and } \mathbf{y}_j,\ i \neq j, \text{ belong to the same class} \\ 0 \quad \text{otherwise.} \end{cases} \tag{14}$$

where $\mathbf{y}_i^{(v)}$ is the i-th sample of the $\mathbf{Y}_v$ from the v-th view and $t^{(v)}$ is the corresponding kernel parameter. The graph Laplacian for each observed data space is obtained as $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{W}^{(v)}$, where $\mathbf{D}^{(v)}$ is a diagonal matrix with $\mathbf{D}_{ii}^{(v)} = \sum_j \mathbf{W}_{ij}^{(v)}$. Because the graph Laplacians from different views vary in their scale, we use the normalized graph Laplacian given by

$$\mathbf{L}_N^{(v)} = \mathbf{D}_v^{-1/2}\mathbf{L}^{(v)}\mathbf{D}_v^{-1/2} = \mathbf{I} - \mathbf{W}_N^{(v)}, \tag{15}$$

where $\mathbf{W}_N^{(v)}$ is the normalized similarity matrix defined as

$$\mathbf{W}_N^{(v)} = \mathbf{D}_v^{-1/2}\mathbf{W}^{(v)}\mathbf{D}_v^{-1/2}. \tag{16}$$

Since the elements of $\mathbf{W}_N^{(v)}$ and $\mathbf{L}_N^{(v)}$ now have the same scale for all views, they can be combined in the joint graph Laplacian as

$$\tilde{\mathbf{L}} = \mathbf{L}_N^{(1)} + \mathbf{L}_N^{(2)} + \ldots + \mathbf{L}_N^{(V)} = \sum_v \mathbf{L}_N^{(v)}, \tag{17}$$

With the graph Laplacian in (17), we define the discriminative shared-space prior as

$$p(\mathbf{X}) = \prod_{v=1}^{V} p(\mathbf{X}|\mathbf{Y}_v)^{\frac{1}{V}} = \frac{1}{V \cdot Z_q} \exp\left[-\frac{\beta}{2}\mathrm{tr}(\mathbf{X}^T\tilde{\mathbf{L}}\mathbf{X})\right], \tag{18}$$

where, as in (10), Z_q is a normalization constant and $\beta > 0$ is a scaling parameter that controls the penalty level incurred by the shared prior. The prior in (18) is the geometric mean of the discriminative priors for each of the target views. As a result, this prior prefers the discriminative shared manifold that maximizes, on average, class-separation of the data from all the views.

3.3 DS-GPLVM: Learning

The learning of the model parameters consists of minimizing the negative log-likelihood subject to the unknown parameters. By combining (12) and (18), we arrive at the following minimization problem

$$\min_{\mathbf{X},\theta_s} L_s = \min_{\mathbf{X},\theta_s} \sum_v L^{(v)} + \frac{\beta}{2}\mathrm{tr}(\mathbf{X}^T\tilde{\mathbf{L}}\mathbf{X}), \tag{19}$$

where $L^{(v)}$ is given by (13) for each view. To minimize L_s, we use the conjugate-gradients algorithm [17] with the gradient of (19) w.r.t. the latent positions $\mathbf{X}$ given by

$$\frac{\partial L_s}{\partial \mathbf{X}} = \sum_v \frac{\partial L^{(v)}}{\partial \mathbf{X}} + \beta \tilde{\mathbf{L}} \mathbf{X}, \tag{20}$$

where we apply the chain rule to the log-likelihood of each view, *i.e.*, $\frac{\partial L^{(v)}}{\partial \mathbf{X}} = \frac{\partial L^{(v)}}{\partial \mathbf{K}_v} \frac{\partial \mathbf{K}_v}{\partial x_{ij}}$, and

$$\frac{\partial L^{(v)}}{\partial \mathbf{K}_v} = \frac{D}{2} \mathbf{K}_v^{-1} - \frac{1}{2} \mathbf{K}_v^{-1} \mathbf{Y}_v \mathbf{Y}_v^T \mathbf{K}_v^{-1}. \tag{21}$$

The gradients of (19) w.r.t. the kernel parameters θ_s are derived in the same way as for the latent positions. The parameters $t^{(v)}$ of the heat kernel in the prior are set using a cross-validation procedure, in order to avoid 'filtering out' the prior by the employed minimization approach. Finally, the weight of the prior β is set using a cross-validation procedure designed to optimize the classification performance of the classifier learned in the shared manifold, as explained in Sec. 4.

3.4 DS-GPLVM: Inference

To perform the inference of a test point from view $v = 1 \dots V$, $\mathbf{y}_i^{(v)}$, we need first to learn inverse mappings from the observation space $\mathbf{Y}_v$ to the shared space $\mathbf{X}$ [23]. This is attained by learning (separately for each view) the following mapping functions

$$x_{ij} = g_j^{(v)}(\mathbf{y}_i^{(v)}; \mathbf{a}) = \sum_{m=1}^{N} a_{jm}^{(v)} k_{bc}^{(v)}(\mathbf{y}_i^{(v)} - \mathbf{y}_m^{(v)}), \tag{22}$$

where x_{ij} is the j-th dimension of $\mathbf{x}_i$, and $g_j^{(v)}$ is modeled using kernel ridge regression over $\mathbf{Y}_v$ for each dimension and each view. To obtain the smooth inverse mapping, we apply the RBF kernel to each dimension of the training data as

$$k_{bc}^{(v)}(\mathbf{y}_i^{(v)} - \mathbf{y}_m^{(v)}) = \exp(-\frac{\gamma_v}{2} \|\mathbf{y}_i^{(v)} - \mathbf{y}_m^{(v)}\|^2), \tag{23}$$

where γ_v are the kernel inverse width parameters for each observation space v. The weight parameters $\mathbf{A}^{(v)}$ of the kernel ridge regression are found in the closed form as

$$\mathbf{A}^{(v)} = \mathbf{X}^T (\mathbf{K}_{bc}^{(v)} + \lambda \mathbf{I})^{-1}, \quad v = 1 \dots V, \tag{24}$$

where $\mathbf{K}_{bc}^{(v)}$ is the kernel matrix computed over the training data from view v. The regularization term $\lambda \mathbf{I}$ helps to stabilize the inverse numerically by bounding the smallest eigenvalues of the kernel matrix away from zero. Note that learning and inference of the models presented in Sec. 2 can be performed following the same procedure with the one explained in this section, using only a single view as input. Finally, once the test sample is projected onto the shared manifold, a classification of the target facial expressions can be accomplished by using different classifiers trained on the shared manifold. In this paper, we employ the linear k-NN classifier.

4 Experiments

We evaluate the performance of the proposed DS-GPLVM on real-world images from the MultiPIE [24] dataset. We use facial images of 270 subjects displaying facial expressions of Neutral (NE), Disgust (DI), Surprise (SU), Smile (SM), Scream (SC) and Squint (SQ) captured at pan angles $-30°$, $-15°$, $0°$, $15°$ and $30°$, resulting in 1500 images per pose. For every image we picked the flash from the view of the corresponding camera in order to always have the same illumination. The images were cropped to have equal size 140×150, and annotations of the locations of 68 facial landmark points were provided by [25], which were used to align the facial images in each pose. From each aligned facial image, we extracted LBPs, with radius 2, resulting in 59 bins. We use LBPs since they have been shown to perform well in facial expression recognition tasks [6]. For the experiments, we used the following three sets of features: (I) facial landmarks (the 68 landmark points), (II) full appearance features (LBPs extracted from the whole face image), and (III) part-based appearance features (LBPs extracted from the facial patches (of size 15×15)) extracted around the facial landmarks.

To reduce the dimensionality of the input features, we applied PCA, resulting in 20 and 70 dimensional inputs for feature sets (I) and (II-III), respectively. Throughout the experiments, we fix the size of the latent space of the tested models to the value for which we obtained the best performance (we used 5D space for the proposed DS-GPLVM). For the kernel methods, we used the RBF kernel with the width parameter set using a validation procedure, as done in [15]. The optimal weight for the prior β was found by another validation procedure, as done in [15]. To report the accuracy of facial expression recognition, we use the classification rate, where the classification is performed using 1-NN classifier for all the tested methods. In all our experiments, we applied 5-fold subject-independent cross-validation procedure.

We compared the DS-GPLVM to the state-of-the-art single- and multi-view methods. The baseline methods include: 1-nearest neighbor (1-NN) classifier trained/tested in the original feature space, LDA, supervised LPP, and their kernel counterparts, the D-GPLVM [15] with the GDA-based prior, and the GPLRF [18]. These are well-known methods for supervised dimensionality reduction applicable to single observation space. We also compared DS-GPLVM to the state-of-the-art methods for multi-view learning, the multi-view extensions of LDA (GMLDA), and LPP (GMLPP) [26].

The evaluation of the tested models is conducted using the data from all poses for training, while testing is performed 'pose-wise', *i.e.*, by using the data from each pose separately. The same strategy was used for evaluation of the multi-view techniques *i.e.*, GMLDA and GMLPP. Table 1 summarises the results for the three sets of features, averaged across the poses. Interestingly, LDA and LPP achieve high performance on the feature set (I). We attribute this to the fact that when points are used as the inputs, sufficiently discriminative pose-wise manifolds can be learned using the linear models. This is because the facial points of different subjects are well aligned, and subject-specific factors, that are present in the texture features, are filtered out. Furthermore, these models outperform (on average) their kernel counterparts (D-GPLVM and GPLRF), and their multi-view extensions (GMLDA and GMLPP) possibly due to the overfitting of these models. Yet, the proposed DS-GPLVM outperforms its 'single-view' counterpart (*i.e.*, GPLRF), which we ascribe to its learning of the shared manifold, that, evidently,

Table 1. Pose-wise FER. Average classification accuracy across all views on MultiPIE database for the three different type of features. DS-GPLVM was trained across all available views and the presented results correspond to back projections from each view to the shared latent space. LDA and LPP are linear models, and they perform well on the facial points. However, they are outperformed by the kernel methods on the appearance features, with the proposed model performing similarly or better than the other kernel-based models. The reported standard deviation is computed from the average results across the views.

Methods	Features		
	I	II	III
kNN	77.22 ± 5.18	61.46 ± 4.09	81.25 ± 2.62
LDA	88.47 ± 8.38	72.28 ± 3.99	85.47 ± 3.07
LPP	88.40 ± 7.99	71.94 ± 4.21	85.51 ± 3.04
D-GPLVM	84.98 ± 5.48	73.64 ± 4.90	84.27 ± 2.43
GPLRF	87.58 ± 5.02	76.89 ± 4.26	86.91 ± 2.81
GMLDA	83.25 ± 6.64	70.89 ± 5.25	84.73 ± 3.09
GMLPP	80.07 ± 3.89	66.28 ± 3.62	82.03 ± 2.45
DS-GPLVM	$\mathbf{88.83 \pm 5.30}$	$\mathbf{77.32 \pm 3.42}$	$\mathbf{87.51 \pm 2.02}$

enhances the classification across all poses. It also performs similarly to the linear models on the feature set (I) but with significantly lower standard deviation, meaning that it achieves more consistent recognition across views. When appearance features are used, learning of the discriminative low-dimensional manifolds is more challenging, as mentioned above. However, the proposed DS-GPLVM achieves similar or better accuracy compared to other single- and multi-view methods due to its successful unraveling of the non-linear manifold shared across different views. Although for these features the results of DS-GPLVM are slightly better than those obtained by GPLRF, the latter learns separate classifiers for each view, in contrast to the DS-GPLVM that uses a single classifier. Note also that the DS-GPLVM retains relatively small variance across poses and feature sets, which makes it more reliable for multi-view recognition.

From Table 1, the feature set (I) achieves slightly better results than feature set (III), however, it is less stable since it results in high standard deviation by all tested models. Considering this, and since we want to test the effectiveness of the proposed model on handling non-linear correlations across the views, we proceed with the experiments on the feature set (III). Table 2 shows the performance of the tested models across all poses for feature set (III). It is evident that in this scenario the proposed DS-GPLVM performs consistently better than the other models across most of the views. It is important to note that although GPLRF slightly outperforms DS-GPLVM in $\pm 30^{\circ}$ pose, the DS-GPLVM significantly outperforms the GPLRF model in the frontal pose, which is the most difficult for expression classification. Again, we attribute this to the fact that DS-GPLVM performs classification in the shared manifold, which, evidently, augments the classification in the frontal pose by using information learned from the other poses.

Finally, we compare on MultiPIE the DS-GPLVM to the sate-of-the-art methods for multi-view facial expression recognition. The results of [4] are obtained from the corresponding paper. To compare our method with [12], we extracted dense SIFT features from the same images we used from MultiPIE. The resulting features were then fed

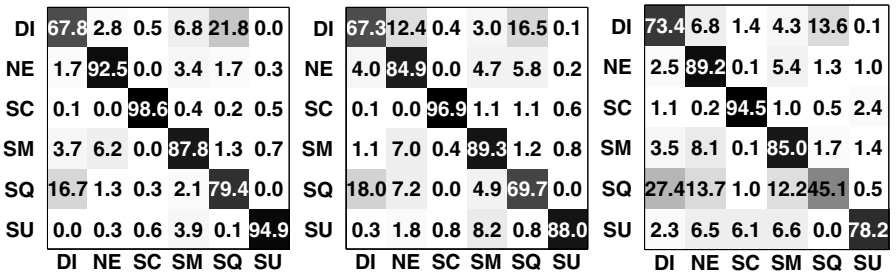
Table 2. Pose-wise FER. Classification accuracy for the MultiPIE dataset across all views for the feature set (III). DS-GPLVM was trained using data from all the views. The results are for the back-projections from each view to the shared latent space. The reported standard deviation is across the 5 folds.

Methods	Poses				
	-30°	-15°	0°	15°	30°
kNN	82.82 ± 0.019	82.43 ± 0.017	76.59 ± 0.034	82.06 ± 0.017	82.37 ± 0.017
LDA	86.62 ± 0.014	87.42 ± 0.015	80.03 ± 0.014	87.11 ± 0.015	86.17 ± 0.012
LPP	86.81 ± 0.014	87.35 ± 0.013	80.09 ± 0.018	86.86 ± 0.017	86.43 ± 0.011
D-GPLVM	84.67 ± 0.017	86.61 ± 0.020	80.36 ± 0.017	85.89 ± 0.019	83.86 ± 0.017
GPLRF	$\mathbf{87.73 \pm 0.026}$	88.87 ± 0.020	81.94 ± 0.025	88.16 ± 0.022	$\mathbf{87.83 \pm 0.025}$
GMLDA	86.03 ± 0.019	86.57 ± 0.016	79.23 ± 0.021	86.16 ± 0.011	85.68 ± 0.018
GMLPP	81.65 ± 0.036	84.61 ± 0.038	78.52 ± 0.034	84.14 ± 0.034	81.25 ± 0.029
DS-GPLVM	87.58 ± 0.008	$\mathbf{89.34 \pm 0.007}$	$\mathbf{84.12 \pm 0.013}$	$\mathbf{89.07 \pm 0.006}$	87.65 ± 0.009

into the SVM classifier, as done in [12]. We also compared our model to [9], where the authors perform pose normalisation of the facial points, which are then classified using the SVM classifier. Table 3 shows comparative results. Note that the methods in [4] and [12] both fail to model correlations between different views, which results either in a huge gap between the accuracy across poses (*e.g.*, [4]) or in a performance bounded by the one achieved in the frontal pose (*e.g.*, [12]). The latter is a product of the sparsification, since the frontal view contains more (redundant) information due to the symmetry of the face. The method in [9] models relations between the poses through the normalization to the frontal pose, however, it achieves significantly better performance in the non-frontal poses after the alignment, an evidence which proves that the used features are more discriminative in the non-frontal views, a fact that was also experienced in [4]. The proposed DS-GPLVM has comparable performance and better than that of the rest of the methods across all the views. Again, we attribute this to the shared manifold, which augments the classification of the under-performing views (mostly in the frontal view). Another worth mentioning fact is that the reported results for our DS-GPLVM are attained using KNN, while for the rest methods we used the linear SVM (a more powerful classifier), as stated in the corresponding papers. The reason we employed KNN is to avoid another cross-validation procedure for parameter tuning. However, our pilot study showed that the performance of the proposed model can be improved by using the SVM.

Table 3. The comparison of tested methods on the MultiPIE database. Our DS-GPLVM, when using the feature set (III), outperforms the state-of-the-art methods for multi-view facial expression recognition. The reported standard deviation is across 5 folds.

Methods	Poses		
	0°	15°	30°
LGBP [4]	82.1	87.3	75.6
Sparse [12]	81.14 ± 0.009	79.25 ± 0.016	77.14 ± 0.019
CGP [9]	80.44 ± 0.017	86.41 ± 0.013	83.73 ± 0.019
DS-GPLVM	$\mathbf{84.12 \pm 0.013}$	$\mathbf{89.07 \pm 0.006}$	$\mathbf{87.65 \pm 0.009}$



(a) DS-GPLVM (b) CGP (c) Sparse

Fig. 2. Comparative confusion matrices for facial expression recognition over all angles of view for the (a) DS-GPLVM, (b) CGP and (c) Sparse

In Fig. 2, we show the confusion matrices for different models trained/tested using the feature set (III). The main source of confusion are the facial expressions of *Disgust* and *Squint*. This is because they are characterized by similar appearance changes in the eyes' region. However, the proposed DS-GPLVM improves significantly the accuracy on *Squint*, compared to the other models. Again, this is because the classification is performed on the shared manifold, which topology is preserved discriminative based on the most informative views.

5 Conclusion

The introduced DS-GPLVM learns a discriminative shared manifold optimized for classification of facial expressions from multiple views. This model is a generalization of existing discriminative latent variable models that learn the manifold of a single observation space. As evidenced by our results on the real data from the MultiPIE dataset, modeling the manifold shared across different views improves 'per-view' classification of facial expressions. Also, the proposed approach outperforms the state-of-the-art methods for supervised multi-view learning, as well as the state-of-the-art methods for multi-view facial expression recognition.

Acknowledgments. This work has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Stefanos Eleftheriadis is further funded in part by the European Community's 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet).

References

1. Pantic, M., Nijholt, A., Pentland, A., Huanag, T.: Human-centred intelligent human? computer interaction (hci^2): how far are we from attaining it? IJAACS 1, 168–187 (2008)
2. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing 27, 1743–1759 (2009)

3. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on PAMI 31, 39–58 (2009)
4. Moore, S., Bowden, R.: Local binary patterns for multi-view facial expression recognition. Computer Vision and Image Understanding 115, 541–558 (2011)
5. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on PAMI 24 (2002)
6. Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., Huang, T.: A study of non-frontal-view facial expressions recognition. In: 19th Int'l Conf. on Pattern Recognition. IEEE (2008)
7. Hesse, N., Gehrig, T., Gao, H., Ekenel, H.K.: Multi-view facial expression recognition using local appearance features. In: 21st Int'l Conf. on Pattern Recognition (ICPR). IEEE (2012)
8. Dornaika, F., Orozco, J.: Real time 3d face and facial feature tracking. Journal of Real-Time Image Processing 2, 35–44 (2007)
9. Rudovic, O., Pantic, M., Patras, I.: Coupled gaussian processes for pose-invariant facial expression recognition. IEEE Transactions on PAMI 35, 1357–1369 (2013)
10. Rudovic, O., Patras, I., Pantic, M.: Regression-based multi-view facial expression recognition. In: Proceedings of Int'l Conf. Pattern Recognition (ICPR 2010), Istanbul, Turkey (2010)
11. Zheng, W., Tang, H., Lin, Z., Huang, T.S.: Emotion recognition from arbitrary view facial images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 490–503. Springer, Heidelberg (2010)
12. Tariq, U., Yang, J., Huang, T.S.: Multi-view facial expression recognition analysis with generic sparse coding feature. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part III. LNCS, vol. 7585, pp. 578–588. Springer, Heidelberg (2012)
13. Lowe, D.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2. IEEE (1999)
14. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Computer Vision and Pattern Recognition. IEEE (2009)
15. Urtasun, R., Darrell, T.: Discriminative gaussian process latent variable model for classification. In: Proc. of the 24th International Conference on Machine Learning. ACM (2007)
16. Lawrence, N.: Probabilistic non-linear principal component analysis with gaussian process latent variable models. The Journal of Machine Learning Research 6, 1783–1816 (2005)
17. Rasmussen, C., Williams, C.: Gaussian processes for machine learning, vol. 1. MIT Press, Cambridge (2006)
18. Zhong, G., Li, W.J., Yeung, D.Y., Hou, X., Liu, C.L.: Gaussian process latent random field. In: Twenty-Fourth AAAI Conference on Artificial Intelligence (2010)
19. Rue, H., Held, L.: Gaussian Markov random fields: theory and applications, vol. 104. Chapman & Hall (2005)
20. Shon, A., Grochow, K., Hertzmann, A., Rao, R.: Learning shared latent structure for image synthesis and robotic imitation. Advances in NIPS 18 (2006)
21. Ek, C., Lawrence, N.: Shared Gaussian Process Latent Variable Models. PhD thesis, Oxford Brookes University (2009)
22. Ek, C.H., Torr, P.H.S., Lawrence, N.D.: Gaussian process latent variable models for human pose estimation. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 132–143. Springer, Heidelberg (2008)
23. Lawrence, N.D., Candela, J.Q.: Local distance preservation in the gp-lvm through back constraints. In: Proc. of the Twenty-Third Int'l Conf. on Machine Learning. ACM (2006)
24. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. IVC 28, 807–813 (2010)
25. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: 5th Workshop on AMFG, Proc. of the Int'l Conf. CVPR-W 2013 (2013)
26. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: IEEE Conference on CVPR (2012)