View-constrained Latent Variable Model for Multi-view Facial Expression Classification

Stefanos Eleftheriadis[1], Ognjen Rudovic[1] and Maja Pantic[1,2]
{s.eleftheriadis, o.rudovic, m.pantic}@imperial.ac.uk

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

Abstract. We propose a view-constrained latent variable model for multi-view facial expression classification. In this model, we first learn a discriminative manifold shared by multiple views of facial expressions, followed by the expression classification in the shared manifold. For learning, we use the expression data from multiple views, however, the inference is performed using the data from a single view. Our experiments on data of posed and spontaneously displayed facial expressions show that the proposed approach outperforms the state-of-the-art methods for multi-view facial expression classification, and several state-of-the-art methods for multi-view learning.

1 Introduction

Facial expression recognition (FER) has been extensively studied in controlled environments where the subjects exhibit posed expressions in a nearly frontal pose [1]. However, in the majority of real-world applications, facial images can be taken from multiple views, depending on the camera position and the people's head movements. For this reason, there is an ever-growing need for automated systems that can accurately perform view-invariant FER[1]. The main challenge here is to decouple rigid facial movements due to the head-pose variation and non-rigid facial movements due to the facial expressions, as these are non-linearly coupled in 2D images [2]. Another challenge is how to effectively exploit the information from multiple views in order to facilitate the FER. Thus, exploiting the fact that each view of a facial expression is a different manifestation of the same underlying facial-expression-related content should lead to simpler and more effective classifiers for the target task.

Recent advances in the field, focus on view-invariant FER, and based on how they deal with the head-pose variations in 2D, they can be classified in those that: (i) perform view-invariant FER ([3–5]), (ii) perform view normalization before FER ([6, 7]), and (iii) learn a single classifier using data from multiple views ([8, 9]). A representative of the first group is [3], where the Local Binary Patterns (LBP) [10] (and its variants) are used to perform a two-step FER: first,

[1] *View-invariant* refers to the case where training is conducted using data from multiple views, and inference is performed with data from a single view.

the view is classified using the Support Vectors Machine (SVM) [11], and then, a view-specific SVM is applied to perform FER. In [4], different appearance features are extracted around the locations of characteristic facial points, and used to train various pose-specific classifiers. Similarly, [5] used per-view-trained 2D Active Appearance Models (AAMs) [12] to locate a set of characteristic facial points, and extract appearance-based features such as Discrete Cosine Transform (DCT) around facial points. By learning separate classifiers for each view, these approaches ignore correlations across the views, which makes them suboptimal for the target task.The approaches in the second group ([6, 7]) first perform view normalization to the frontal through Coupled Gaussian Process (CGP) regression, and then apply a classifier trained in the frontal view. A limitation of this approach is that the view normalization and learning of the expression classifier are done independently, thus bounding the performance of FER by the accuracy of the view normalization.The third group ([8, 9]) uses a single classifier learned using the expression data from multiple views. For instance, [8] used variants of dense Scale Invariant Feature Transform (SIFT) [13] features extracted from multi-view facial expression images. Likewise, [9] used the Generic Sparse Coding scheme ([14]) to learn a dictionary of SIFT features extracted from facial images in different views.

However, while these methods focus on the feature extraction step, they fail to model relationships between the views in a principled manner, resulting in view-specific classifiers, or a single classifier with high complexity for large number of views/expressions. To our knowledge, the only method proposed so far that addresses these limitations is the recently proposed Discriminative Shared Gaussian Process Latent Variable Model (DS-GPLVM) [15]. In this method, the authors use the notion of Shared Gaussian Processes (GP) [16] to learn a manifold shared among multiple views of facial expressions, where the discriminative shared-space prior is placed over the manifold. Then, the view-invariant expression classification is performed during inference using a classifier learned in the manifold. However, a bottleneck of this approach is that the back-mappings, i.e., projections of the data observed in different views to the manifold, are learned *independently* of the manifold. This can lead to the overfitting of the mappings, making it difficult for the model to generalize to novel data. This, in turn, can adversely affect facial expression classification in the shared manifold.

To address this, we propose the View-constrained Gaussian Process Model (VC-GPM) that generalizes the DS-GPLVM mentioned above by learning the structure of the shared manifold and the back-mappings simultaneously. In this way, we further constrain the structure of the manifold, allowing the model to find more accurate mappings from the observed data to the manifold, and, thus, improve their classification. For this, we propose a novel learning algorithm based on the Alternating Direction Method (ADM) [17] and *leave-one-out* learning strategy. Specifically, we split the learning into different sub-problems (for each view and the manifold), where the optimization of each sub-problem is done separately, making it computationally tractable for a large number of views. Also, instead of using the standard kernel regression to learn the back-mappings (as done in DS-GPLVM), we formulate *leave-one-out* learning of the mappings that results in the learned back-mappings being more robust to subject differences and other expression-unrelated sources of variation in the observed views of facial expressions. We show in our experiments on data of posed and spontaneously displayed facial expressions that the proposed VC-GPM outperforms the state-of-the-art methods for view-invariant FER as well as several state-of-the-art multi-view learning methods.

2 View-constrained GP Model (VC-GPM)

2.1 VC-GPM: Model Formulation

Let us assume we are given a set of V views, $\mathbf{Y} = \{\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(V)}\}$, where each view is represented with a high-dimensional observation space $\mathbf{Y}^{(v)} = [\mathbf{y}_1^{(v)}, \dots, \mathbf{y}_N^{(v)}]^T \in \mathcal{R}^{N \times D}$, $v = 1 \dots V$, and N, D are the number of data samples and the size of the observation space, respectively. We seek to find a low-dimensional shared manifold $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T \in \mathcal{R}^{N \times q}$, where $q \ll D$ is the size of the manifold that generates all V views simultaneously. Formally, using the Shared GPs [16, 18] framework, we can write the joint likelihood of V views as

$$p(\mathbf{Y}|\mathbf{X}, \theta_s) = p(\mathbf{Y}^{(1)}|\mathbf{X}, \theta^{(1)}) \times \dots \times p(\mathbf{Y}^{(V)}|\mathbf{X}, \theta^{(V)}), \tag{1}$$

where the likelihood of the observed data from the view v, given the manifold, is

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

Here, $\mathbf{K}^{(v)}$ is the kernel matrix, and its elements are obtained by applying the covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$, to each data pair. The covariance function is usually chosen as the sum of the Radial Basis Function (RBF), bias and noise term, i.e.,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp(-\frac{\theta_2}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2) + \theta_3 + \frac{\delta_{i,j}}{\theta_4}, \tag{3}$$

where $\delta_{i,j}$ is the Kronecker delta function, and $\theta^{(v)} = (\theta_1^{(v)}, \theta_2^{(v)}, \theta_3^{(v)}, \theta_4^{(v)})$ are the kernel parameters [19]. It is further assumed that each observation space is generated from the shared manifold via a separate GP, governed by the parameters stored in $\theta_s = \{\theta^{(1)}, \dots, \theta^{(V)}\}$. To find the shared manifold $\mathbf{X}$ that is discriminative, we compute the posterior distribution $p(\mathbf{X}, \theta_s|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \theta_s) p(\mathbf{X})$. This allows us to include our prior knowledge about the class information (i.e., the expression category) into the learning task, as described below.

Discriminative shared-space prior. To define a discriminative shared-space prior for multi-view learning, we use the graph Laplacian matrix. To this end, we first construct the view-specific weight matrices $\mathbf{W}^{(v)}$, by accounting for the data location along with the class. Specifically, the elements of $\mathbf{W}^{(v)}$ are obtained from the RBF kernel on each view

$$\mathbf{W}_{ij}^{(v)} = \begin{cases} \exp\left(-\frac{\|\mathbf{y}_i^{(v)} - \mathbf{y}_j^{(v)}\|^2}{t^{(v)}}\right) & \text{if } i \neq j \text{ and } c_i = c_j, \\ 0 \quad \text{otherwise.} \end{cases} \tag{4}$$

with $\mathbf{y}_i^{(v)}$ the i-th sample in $\mathbf{Y}^{(v)}$, c_i the label, and $t^{(v)}$ the kernel width which is set to the mean squared distance of the data. Then, the Laplacian for view v is $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{W}^{(v)}$, where $\mathbf{D}_{ii}^{(v)} = \sum_j \mathbf{W}_{ij}^{(v)}$. Because the Laplacians of different views vary in their scale, we normalize them as $\mathbf{L}_N^{(v)} = (\mathbf{D}^{(v)})^{-1/2}\mathbf{L}^{(v)}(\mathbf{D}^{(v)})^{-1/2}$. Consequently, the joint (regularized) Laplacian is defined as

$$\tilde{\mathbf{L}} = \mathbf{L}_N^{(1)} + \mathbf{L}_N^{(2)} + \ldots + \mathbf{L}_N^{(V)} + \xi\mathbf{I} = \sum_v \mathbf{L}_N^{(v)} + \xi\mathbf{I}, \tag{5}$$

with $\mathbf{I}$ the identity matrix, and ξ a parameter, which ensures that $\tilde{\mathbf{L}}$ is positive-definite. This, allows us to define the discriminative shared-space prior as

$$p(\mathbf{X}) = \prod_{v=1}^{V} p(\mathbf{X}|\mathbf{Y}^{(v)})^{\frac{1}{V}} = \frac{1}{V \cdot Z_q} \exp\left[-\frac{\beta}{2}\mathrm{tr}(\mathbf{X}^T\tilde{\mathbf{L}}\mathbf{X})\right]. \tag{6}$$

Here, Z_q is a normalization constant and $\beta > 0$ is a scaling parameter. The shared-space prior in (6) aims at maximizing the class separation in the manifold learned from data from all the views. Using this prior and Eq. (8), the negative log-likelihood of the proposed model is given by

$$L_s(\mathbf{X}) = \sum_v L^{(v)} + \frac{\beta}{2}\mathrm{tr}(\mathbf{X}^T\tilde{\mathbf{L}}\mathbf{X}), \tag{7}$$

where $L^{(v)}$ is the negative log-likelihood of data from view v, and is given by

$$L^{(v)} = \frac{D}{2}\ln|\mathbf{K}^{(v)}| + \frac{1}{2}\mathrm{tr}[(\mathbf{K}^{(v)})^{-1}\mathbf{Y}^{(v)}(\mathbf{Y}^{(v)})^T] + \frac{ND}{2}\ln 2\pi. \tag{8}$$

Finally, the shared manifold $\mathbf{X}$ and the kernel parameters θ_s are found by minimizing the negative log-likelihood in Eq.(7) (e.g., using conjugate gradients algorithm [19]).

View constraints. The model described above finds the data manifold shared among multiple views, however, to embed newly observed data in this manifold, we need to learn the back-mappings from the observed space/s to the manifold. Another role of these back-mappings is to constrain the learning of the shared manifold by acting as additional regularizers in the model, enforcing the data that are close in the observation space to be close on the manifold. This cannot be attained by the discriminative prior introduced above as it ensures the opposite - that the data close on the manifold are close in the observation space. Therefore, we define V sets of constraints that enforce separate inverse mappings from each view to the shared space. We refer to these as independent back-projections (IBP), and they are given by

$$\underbrace{\mathbf{X} = g(\mathbf{Y}^{(v)}, \mathbf{A}^{(v)}) = \mathbf{K}_{bc}^{(v)}\mathbf{A}^{(v)}}_{\textbf{IBP}\text{ from each view } v=1,\ldots,V} \tag{9}$$

where $g(\cdot,\cdot)$ represents the mapping functions. The elements of $\mathbf{K}_{bc}^{(v)}$ are given by $k_{bc}(\mathbf{y}_i, \mathbf{y}_m) = \exp(-\frac{\gamma}{2}\|\mathbf{y}_i - \mathbf{y}_m\|^2)$ with γ being the inverse width of the kernel. Note that for a single view, the model can be re-parameterized to obtain

$\mathbf{X}$ as a function of the back-mapping parameters (see [20]). Yet, for multiple views, this is not feasible as it would result in different $\mathbf{X}$ for each view. In what follows, we incorporate the view constraints defined above into the objective function of the VC-GPM given by Eq.(7), and propose a novel algorithm for the model learning and inference. In this way, we overcome the limitations of the DS-GPLVM [15], which fails to account for the view-constraints while learning the shared manifold.

2.2 VC-GPM: Learning and Inference

Learning of the model parameters consists of minimizing the negative log-likelihood given by Eq. (7) subject to the IBP constraints:

$$\min_{\mathbf{X},\theta_s,\mathbf{A}} L_s(\mathbf{X}) + R(g) \quad s.t. \quad IBP(\mathbf{X}, \mathbf{A}^{(v)}) \triangleq \mathbf{X} - \mathbf{K}_{bc}^{(v)}\mathbf{A}^{(v)} = \mathbf{0} \ , v = 1, \ldots, V \tag{10}$$

where $R(g)$ is the regularizer defined in the space of $g(\cdot,\cdot)$. By applying Representer Theorem, we obtain the optimal functional form for $R(g)$, given by

$$R(g) = \sum \frac{\lambda^{(v)}}{2} r(g^{(v)}), \ r(g^{(v)}) = \ \mathrm{tr}((\mathbf{A}^{(v)})^T \mathbf{K}_{bc}^{(v)} \mathbf{A}^{(v)}). \tag{11}$$

Parameter Optimization. Herein, we present the learning procedure for the IBP-constrained model. From Eq. (10), we see that the back-mapping from each view is represented by an independent set of linear constraints. We exploit this to find the model parameters by iteratively solving a set of sub-problems. To this end, we use the Lagrange multipliers to incorporate the IBP constraints into the regularized log-likelihood of Eq. (10) resulting in the Augmented Lagrangian (AL) function:

$$\begin{aligned} &\mathcal{L}^{IBP}(\mathbf{X}, \{\mathbf{A}^{(v)}, \mathbf{\Lambda}^{(v)}\}_{v=1}^{V}) = L_s(\mathbf{X}) + R(g) + \\ &\sum_{v=1}^{V} \langle \mathbf{\Lambda}^{(v)}, IBP(\mathbf{X}, \mathbf{A}^{(v)}) \rangle + \frac{\mu}{2} \sum_{v=1}^{V} \| IBP(\mathbf{X}, \mathbf{A}^{(v)}) \|_F^2, \end{aligned} \tag{12}$$

with $\mathbf{\Lambda}^{(v)}$ the Lagrange multiplier for view v, $\langle\cdot,\cdot\rangle$ the inner product, and μ a penalty parameter. The objective function in Eq. (12) is separable, and hence, we employ the Alternating Direction Method (ADM) [17] to decompose it into subproblems. Specifically, we split the learning of the shared space and the back-mappings from each view, following the iterations of ADM: first solve for $\mathbf{X}, \theta_s$

$$\{\mathbf{X}, \theta_s\}_{t+1} = \arg\min_{\mathbf{X},\theta_s} L_s(\mathbf{X}) \ + \frac{\mu_t}{2} \sum_{v=1}^{V} \| IBP(\mathbf{X}, \mathbf{A}_t^{(v)}) + \frac{\mathbf{\Lambda}_t^{(v)}}{\mu_t} \|_F^2. \tag{13}$$

Then, for each view $v = 1, ..., V$, we solve for $\mathbf{A}^{(v)}$ as

$$\mathbf{A}_{t+1}^{(v)} = \arg\min_{\mathbf{A}^{(v)}} r(\mathbf{A}^{(v)}) \ + \frac{\mu_t}{2} \| IBP(\mathbf{X}_{t+1}, \mathbf{A}^{(v)}) + \frac{\mathbf{\Lambda}_t^{(v)}}{\mu_t} \|_F^2, \tag{14}$$

and finally update the Lagrangian and the penalty parameter as

$$\mathbf{\Lambda}_{t+1}^{(v)} = \mathbf{\Lambda}_t^{(v)} + \mu_t IBP(\mathbf{X}_{t+1}, \mathbf{A}_{t+1}^{(v)}) \quad \text{and} \quad \mu_{t+1} = \min(\mu_{max}, \rho\mu_t), \tag{15}$$

respectively. Note that in Eq. (15), ρ is constant (it is typically set to $\rho = 1.1$).

Since there is not a closed-form solution for the problem in Eq. (13), we use the conjugate gradient algorithm (CG) [19] to minimize the objective w.r.t. $\mathbf{X}$ and θ_s. On the other hand, the problem in Eq. (14) is similar to that of the regularized Kernel Ridge Regression (KRR) [21] and has a closed-form solution:

$$\mathbf{A}^{(v)} = (\mathbf{K}_{bc}^{(v)} + \frac{\lambda^{(v)}}{\mu_t}\mathbf{I})^{-1}(\mathbf{X} + \frac{\mathbf{\Lambda}_t^{(v)}}{\mu_t}). \tag{16}$$

However, this solution depends on the parameters $\gamma^{(v)}$ and $\lambda^{(v)}$, which need to be tuned through costly cross-validation procedures. To alleviate this, we use the notion of the Leave-One-Out (LOO) cross-validation procedure for the KRR [21] to define learning of $\gamma^{(v)}$ and $\lambda^{(v)}$, and, thus, obtain $\mathbf{A}^{(v)}$ indirectly.

The learning in LOO is based on the fact that given any training set and the corresponding regression model, if we add a sample to the training set with the target equal to the output predicted by the model, the latter will not change since the cost will not increase [21]. Thus, given the training set with the sample $\mathbf{y}_i^{(v)}$ left out, the predicted outputs $\hat{\mathbf{X}}^{(-i)}$ [2] will not change if the sample $\mathbf{y}_i^{(v)}$ with target $\hat{\mathbf{x}}_i^{(-i)}$ is added to the set. Then, the goal of LOO is to minimize the difference between the prediction $\hat{\mathbf{x}}_i^{(-i)}$ and the actual output $\mathbf{x}_i$ for all samples. For this, we first define the matrix

$$\mathbf{M} \triangleq \begin{bmatrix} m_{ii} & \mathbf{m}_i^T \\ \mathbf{m}_i & \mathbf{M}_i \end{bmatrix} = (\mathbf{K}_{bc}^{(v)} + \frac{\lambda^{(v)}}{\mu_t}\mathbf{I}), \tag{17}$$

where we partitioned the inverse matrix from Eq. (16) so that the elements corresponding to the i-th sample appear only in the first row and column of $\mathbf{M}$ ($\mathbf{X}$ and $\mathbf{\Lambda}_t^{(v)}$ are also reordered to have the i-th row on the top). We also denote with $\mathbf{M}_i = (\mathbf{K}_{bc\backslash i}^{(v)} + \frac{\lambda^{(v)}}{\mu_t}\mathbf{I}_{N-1})$ the kernel matrix formed from the remaining elements. Then, using Eq. (16), the prediction and the actual target for sample i are

$$\hat{\mathbf{x}}_i^{(-i)} = \mathbf{m}_i^T\mathbf{M}_i^{-1}\mathbf{m}_i\mathbf{A}_i^{(v)} + \mathbf{m}_i^T\mathbf{A}_{-i}^{(v)}, \quad \mathbf{x}_i = m_{ii}\mathbf{A}_i^{(v)} + \mathbf{m}_i^T\mathbf{A}_{-i}^{(v)} - \mathbf{\Lambda}_i^{(v)}/\mu_t. \tag{18}$$

We can now define the cost for the LOO procedure, which is

$$E_{LOO} = \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i^{(-i)}\|^2 = \frac{1}{2}\sum_{i=1}^{N} \|\frac{\mathbf{A}_i^{(v)}}{[\mathbf{M}^{-1}]_{ii}} - \frac{\mathbf{\Lambda}_i^{(v)}}{\mu_t}\|^2 \tag{19}$$

We minimize E_{LOO} w.r.t. $\gamma^{(v)}$ and $\lambda^{(v)}$ using again CG, and then obtain $\mathbf{A}^{(v)}$ from Eq. (16). By adopting the LOO approach, we: (i) avoid the burden of cross-validation, and (ii) reduce the chances of overfitting the model parameters.

[2] The superscript denotes that the i-th sample is out

Algorithm 1 VC-GPM: Learning and Inference

```
Learning
Inputs: $\mathcal{D} = (\mathbf{Y}^{(v)}, \mathbf{c}), v = 1, \ldots, V$. Initialize $\mu_{max} \gg \mu_0 > 0$, $\rho = const.$, $\mathbf{X}_0$, $\mathbf{A}_0^{(v)}$, $\mathbf{\Lambda}_0^{(v)}$.
repeat
    Step 1: Update $(\mathbf{X}, \theta_s)$ by minimizing Eq. (13).
    Step 2: Minimize $E_{LOO}$ from Eq. (19) w.r.t $(\gamma^{(v)},\lambda^{(v)})_{v=1,\ldots,V}$
    Step 3: Update $(\mathbf{\Lambda}^{(v)}$, $\mu$, $\mathbf{A}^{(v)})$ from Eq. (15)-(16).
until convergence of Eq. (12)
Outputs: $\mathbf{X}$, $\mathbf{A}$

Inference
Inputs: $\mathbf{y}^{(v)*}$
Step 1: Find the projection $\mathbf{x}^*$ to the latent space using Eq. (9)
Step 2: Apply kNN classifier to the latent space to obtain the class prediction: $c^* = \text{kNN}(\mathbf{x}^*, \mathbf{X})$.
Output: $c^*$
```

Inference in VC-GPM is straightforward. Test data $\mathbf{y}^*$ are first projected to the shared space using the back-mappings from Eq. (9). Then, classification of the target facial expression is accomplished by using a single classifier (we used the k-NN classifier) trained directly in the learned shared space. Alg.1 summarizes the learning and inference of the proposed VC-GPM.

3 Experiments

3.1 Datasets and Experimental Procedure

We evaluate the performance of the proposed VC-GPM on two publicly available datasets: MultiPIE [22] and Labeled Face Parts in the Wild (LFPW) [23]. From MultiPIE we used images of 270 subjects depicting acted facial expressions of Neutral, Disgust, Surprise, Smile, Scream and Squint, captured at pan angles $-30°$, $-15°$, $0°$, $15°$ and $30°$, resulting in 1531 images per pose. For all images, we selected the flash from the view of the corresponding camera in order to have the same illumination. The LFPW dataset contains images downloaded from google.com, flickr.com, and yahoo.com, depicting spontaneous facial expressions, in large variation of poses, illumination and occlusion. We used 200 images of Neutral and Smile expressions from the test set provided by [23] and manually annotated them in terms of the poses used in MultiPIE. All images were cropped to 140×150 pixels, and the 68 facial landmark points provided by [24] were used to align the facial images in each pose. For the experiments on MultiPIE, we used three feature sets: (I) facial points, (II) LBPs, and (III) DCT. The LBPs and DCT were extracted from local patches of size 15×15 around the facial landmarks. For LBPs, we used 8 neighbors with radius 2, and in DCT we kept the first 15 coefficients of each patch. Note that LBP and DCT are complementary

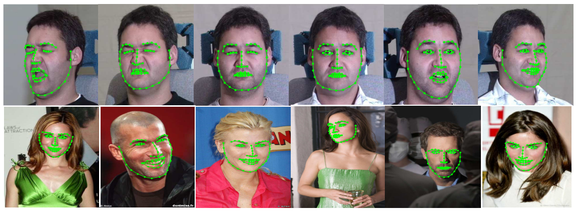

Fig. 1: Example images from the MultiPIE dataset (top) and the LFPW dataset (bottom).

features, since the former captures local information between neighborhood of pixels, while the latter preserves the spatial correlation. On LFPW, we used only feature set (I). To reduce the dimensionality, we applied PCA (95% of variance).

We compared the VC-GPM to the state-of-the-art single- and multi-view learning methods for view-invariant FER. As the baseline method, we use the 1-nearest neighbor (1-NN) classifier in the original feature space. Similarly, we apply 1-NN classifier to the subspace obtained by LDA [25], supervised LPP [26], D-GPLVM [20] and GPLRF [27]. We also compared VC-GPM to several state-of-the-art methods for multi-view learning, namely, the Multi-view Discriminant Analysis (mvDA) [28], and methods for Generalized Multiview Analysis (GMA) [29], namely, GM Linear Discriminant Analysis (GMLDA) and GM Locality Preserving Projections (GMLPP), which extend the LDA and LPP [30] to multiple views. Lastly, we also include the results obtained by DS-GPLVM [15], with the GP kernel parameters and the discriminative prior as in our VC-GPM. However, learning of the shared manifold and the back-mappings in DS-GPLVM is done independently. In all kernel methods, we used the RBF. The width of the kernel, as well as the optimal weight for the prior β (for GPLVM-based models) were set using a cross validation as in [20]. In all the manifold-based methods, we set the size of the manifolds to 5, as it performed best on average.

3.2 Comparisons with Multi-view Learning Methods on MultiPIE

We evaluate the proposed VC-GPM model across views in view-invariant FER. All multi-view learning methods were tested using the same setting. The single-view methods were trained/tested per view and by concatenating features from multiple views. Table 1 summarizes the results for the three feature sets, on MultiPIE. We see that the facial points result in a more discriminative descriptor for all methods. Evidently, VC-GPM outperforms the other models on all three feature sets, showing that it can successfully unravel the discriminative view-constrained space that is better suited for FER. Interestingly, LDA- and LPP-based methods achieve high accuracy, which is comparable to that of D-GPLVM and GPLRF. Moreover, GMLDA and GMLPP perform similarly to their single view trained counterparts for view-invariant FER, indicating that they were not able to fully benefit from the presence of additional views. We also observe a similar performance of the MvDA and the standard LDA. Note also that the accuracy of VC-GPM is higher for 3% than that of GPLRF, which

Table 1: Average classification rate on MultiPIE. The standard deviation is across five views.

Methods	Feature Set		
	I	II	III
kNN	76.15 ± 5.4	81.71 ± 2.9	71.80 ± 2.2
LDA	87.72 ± 6.7	86.24 ± 2.3	87.02 ± 2.6
LPP	87.81 ± 6.7	86.16 ± 2.2	86.82 ± 2.6
D-GPLVM	87.17 ± 5.8	85.92 ± 2.9	86.87 ± 3.2
GPLRF	86.93 ± 6.3	85.58 ± 2.7	86.88 ± 2.9
GMLDA	86.72 ± 6.6	85.18 ± 2.9	86.40 ± 3.4
GMLPP	87.74 ± 6.1	86.10 ± 2.1	86.21 ± 2.1
MvDA	87.84 ± 6.5	86.66 ± 2.8	86.79 ± 2.9
DS-GPLVM	88.64 ± 5.6	87.13 ± 2.7	87.34 ± 2.9
VC-GPM	**90.60 ± 5.4**	**88.44 ± 2.8**	**89.18 ± 2.8**

Table 2: Classification rate on MultiPIE for the feature set (I), *i.e.* the facial points.

Methods	Poses				
	−30°	−15°	0°	15°	30°
GPLRF	91.65 ± 0.017	93.77 ± 0.007	77.59 ± 0.021	85.66 ± 0.026	86.01 ± 0.008
GMLDA	90.47 ± 0.012	94.18 ± 0.007	76.60 ± 0.029	86.64 ± 0.032	85.72 ± 0.015
GMLPP	91.86 ± 0.013	94.13 ± 0.002	78.16 ± 0.013	87.22 ± 0.023	87.36 ± 0.008
MvDA	92.49 ± 0.011	94.22 ± 0.014	77.51 ± 0.022	87.10 ± 0.031	87.89 ± 0.010
DS-GPLVM	92.25 ± 0.013	94.83 ± 0.014	80.18 ± 0.025	87.63 ± 0.017	88.32 ± 0.023
VC-GPM	**93.55 ± 0.019**	**96.96 ± 0.012**	**82.42 ± 0.018**	**89.97 ± 0.023**	**90.11 ± 0.028**

is a special case of VC-GPM. We attribute this to the ability of VC-GPM to integrate the discriminative information from multiple views into the shared space. It also outperforms DS-GPLVM, which fails to account for the view-constraints.

We derive similar conclusions from Table 2, which shows the results by the models tested per view using the best performing feature set, *i.e.* feature set (I). Note that the proposed VC-GPM improves the accuracy in the frontal view significantly, contrary to the rest of the models. Furthermore, it is worth noting that the models' accuracy on the negative pan angles (left side of the face) is higher than on the corresponding positive pan angles (right side of the face). This is in agreement with the recent findings in [31] that show that the left hemisphere of the face is more informative when it comes to expressing negative emotions (*e.g.*, Disgust) since MultiPIE contains more examples of the negative emotion expression. On the other hand, the right hemisphere is more informative for positive emotions (*e.g.*, Happiness).

3.3 Comparisons with other Multi-view Methods

We also compared the VC-GPM (using feature set (III)) with the state-of-the-art methods for view-invariant FER. The results for the LGBP-based method are obtained from [3]. For the method in [9], we extracted Sparse SIFT (SSIFT) features from the same images we used from MultiPIE. In both of the aforementioned methods, the features are fed into the view-specific (for LGBP) or universal (SSIFT) SVM classifier. For the Coupled GP (CGP) [6], first view-normalization is performed by projecting the facial points (feature set (I)) from non-frontal views to the 15° view, as it turns out to be the most discriminative

Table 3: Classification rate on MultiPIE and LFPW, for FER and smile detection respectively.

(a) FER on MultiPIE

Methods	Poses		
	0°	15°	30°
LGBP [3]	82.1	87.3	75.6
SSIFT [9]	81.14	79.25	77.14
CGP [6]	80.44	86.41	83.73
DS-GPLVM	83.73	88.41	87.69
VC-GPM	**84.31**	**89.21**	**90.26**

(b) Smile detection on LFPW (cross dataset)

Method	Poses				
	−30°	−15°	0°	15°	30°
GMLDA	69.00	43.00	80.94	55.76	76.00
GMLPP	**70.00**	47.50	81.25	57.58	79.66
MvDA	**70.00**	50.00	81.25	51.52	**80.00**
DS-GPLVM	57.20	52.50	84.00	69.38	**80.00**
VC-GPM	55.33	**58.00**	**90.00**	**74.55**	**80.00**

for FER. Then, FER is performed by applying the SVM to the pose-normalized features. From Table 3(a), we observe that VC-GPM outperforms, on average, the appearance-based methods. This difference is in part due to the features used and in part due to the fact that the methods in [3] and [9] both fail to model correlations between different views. By contrast, the CGP method accounts for the relations between the views in a pair-wise manner, while VC-GPM does so for all the views simultaneously. However, the proposed VC-GPM shows superior performance to that of DS-GPLVM, which, in turn, outperforms CGP.

3.4 Cross Dataset Experiments on MultiPIE and LFPW

Finally, we test the ability of VC-GPM to generalize to unseen real-world spontaneous data. To this end, we evaluate the models trained on MultiPIE and tested on LFPW, on the smile detection task using feature set (I). This is a challenging task since the test images are captured in uncontrolled environment, with large variation in illumination and occlusions. Also, the models are trained using data of *posed* expressions, which can differ considerably in subtlety compared to the *spontaneous* expressions used for testing. The difficulty of the task is evidenced by the results in Table 3(b), where we observe a significant drop in accuracy of all methods. We also observe that the positive degrees are most informative for smile detection (for the reasons explained above). However, all methods attain higher accuracy in the frontal pose. We attribute this to the fact that because the test images belong to a continuous range from 0° to ±30°, inaccuracies in the pose registration adversely affected the performance of the models. Nevertheless, the proposed VC-GPM outperforms the rest by a large margin in all poses except −30°. We inspected the number of test examples of smiles in this pose, and found that only few were available. Therefore, this misclassification caused a significant drop in the performance of both VC-GPM and DS-GPLVM.

4 Conclusion

We proposed the View-constrained Latent Variable Model for classification of facial expressions from multiple views. Compared to the DS-GPLVM that learns the shared manifold separately from the back-mappings, we showed that constraining the manifold by the proposed per-view constraints results in a more effective model for the target task. As evidenced by our results on the data of posed and spontaneously displayed facial expressions, the proposed model improves per-view FER, compared to that attained by the state-of-the-art methods for supervised multi-view learning and FER.

Acknowledgments

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

References

1. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. on PAMI **31** (2009) 39–58
2. Zhu, Z., Ji, Q.: Robust real-time face pose and facial expression recovery. In: IEEE Int'l Conf. on CVPR. (2006) 681–688
3. Moore, S., Bowden, R.: Local binary patterns for multi-view facial expression recognition. CVIU **115** (2011) 541–558
4. Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., Huang, T.: A study of non-frontal-view facial expressions recognition. In: ICPR. (2008) 1–4
5. Hesse, N., Gehrig, T., Gao, H., Ekenel, H.K.: Multi-view facial expression recognition using local appearance features. In: ICPR. (2012) 3533–3536
6. Rudovic, O., Pantic, M., Patras, I.: Coupled gaussian processes for pose-invariant facial expression recognition. IEEE Trans. on PAMI **35** (2013) 1357–1369
7. Rudovic, O., Patras, I., Pantic, M.: Regression-based multi-view facial expression recognition. In: Proc. of ICPR. (2010) 4121–4124
8. Zheng, W., Tang, H., Lin, Z., Huang, T.: Emotion recognition from arbitrary view facial images. ECCV (2010) 490–503
9. Tariq, U., Yang, J., Huang, T.: Multi-view facial expression recognition analysis with generic sparse coding feature. In: ECCV-W'12. (2012) 578–588
10. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on PAMI **24** (2002) 971–987
11. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning (1995) 273–297
12. Cootes, T.F., Edwards, G.J., Taylor, C.J., et al.: Active appearance models. IEEE Trans. on PAMI **23** (2001) 681–685
13. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV. Volume 2. (1999) 1150–1157
14. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conf. on CVPR. (2009) 1794–1801
15. Eleftheriadis, S., Rudovic, O., Pantic, M.: Shared gaussian process latent variable model for multi-view facial expression recognition. In: ISVC. (2013) 527–538
16. Shon, A., Grochow, K., Hertzmann, A., Rao, R.: Learning shared latent structure for image synthesis and robotic imitation. NIPS **18** (2006) 1233

17. Bertsekas, D.P.: Constrained optimization and lagrange multiplier methods. Computer Science and Applied Mathematics, Boston: Academic Press, 1982 **1** (1982)
18. Ek, C., Lawrence, P.: Shared Gaussian Process Latent Variable Models. PhD thesis, Oxford Brookes University (2009)
19. Rasmussen, C., Williams, C.: Gaussian processes for machine learning. Volume 1. MIT press Cambridge, MA (2006)
20. Urtasun, R., Darrell, T.: Discriminative gaussian process latent variable model for classification. In: ICML, ACM (2007) 927–934
21. Sundararajan, S., Keerthi, S.S.: Predictive approaches for choosing hyperparameters in gaussian processes. Neural Computation **13** (2001) 1103–1118
22. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28** (2010) 807–813
23. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Conf. on CVPR. (2011) 545–552
24. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: CVPR-W13. (2013)
25. Bishop, C.: Pattern recognition and machine learning. Volume 4. Springer (2006)
26. Zheng, Z., Yang, F., Tan, W., Jia, J., Yang, J.: Gabor feature-based face recognition using supervised locality preserving projection. Signal Processing **87** (2007)
27. Zhong, G., Li, W.J., Yeung, D.Y., Hou, X., Liu, C.L.: Gaussian process latent random field. In: AAAI Conference on Artificial Intelligence. (2010)
28. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.: Multi-view discriminant analysis. In: ECCV. (2012) 808–821
29. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: IEEE Conf. on CVPR. (2012) 2160–2167
30. Niyogi, X.: Locality preserving projections. In: NIPS. Volume 16. (2004) 153
31. Pantic, M., Patras, I.: Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans. on SMCB - Part B **36** (2006) 433–449