Joint Facial Action Unit Detection and Feature Fusion: A Multi-Conditional Learning Approach

Stefanos Eleftheriadis, Ognjen Rudovic, *Member, IEEE*, and Maja Pantic, *Fellow, IEEE*

***Abstract*— Automated analysis of facial expressions can benefit many domains, from marketing to clinical diagnosis of neurodevelopmental disorders. Facial expressions are typically encoded as a combination of facial muscle activations, *i.e.*, action units. Depending on context, these action units co-occur in specific patterns, and rarely in isolation. Yet, most existing methods for automatic action unit detection fail to exploit dependencies among them, and the corresponding facial features. To address this, we propose a novel multi-conditional latent variable model for simultaneous fusion of facial features and joint action unit detection. In particular, the proposed model performs feature fusion in a generative fashion via a low-dimensional shared subspace, while simultaneously performing action unit detection using a discriminative classification approach. We show that by combining the merits of both approaches, the proposed methodology outperforms existing purely discriminative/generative methods for the target task. To reduce the number of parameters, and avoid overfitting, a novel Bayesian learning approach based on Monte Carlo sampling is proposed, to integrate out the shared subspace. We validate the proposed method on posed and spontaneous data from three publicly available data sets (CK+, DISFA, and Shoulder-pain), and show that both feature fusion and joint learning of action units leads to improved performance compared with the state-of-the-art methods for the task.**

***Index Terms*— Multiple action unit detection, multi-conditional learning, multi-label, Gaussian processes.**

I. Introduction

FACIAL expression is one of the most powerful channels of non-verbal communication [1]. It conveys emotions, provides clues about people's personality and intentions, reveals the state of pain, weakness or hesitation, among others. Automatic analysis of facial expressions has attracted significant research attention over the past decade, due to its wide importance in various domains such as medicine, security and psychology [2]. The facial action coding system (FACS) [3] is the most comprehensive anatomically-based system for describing facial expressions in terms of non-overlapping, visually detectable facial muscle activations, named action units (AUs). FACS defines 32 unique AUs, several categories of head/eye positions and other movements, which can be used to describe every possible facial expression. Automatic detection of AUs is a challenging task mainly due to the complexity and subtlety of human facial behavior, but also due to individual differences variations in head-pose, illumination, occlusions, etc [2]. In computer vision, these sources of variation in facial expression data are typically accounted for at (i) the feature level, by deriving facial features that are robust to the aforementioned variations, and/or (ii) the model level, by capturing temporal dynamics of AUs (e.g., changes in AU intensity over time) and semantics of AUs, *i.e.*, their co-occurrences, as commonly encountered in spontaneous data.

At the feature level, detection of AUs can be performed using either geometric or appearance features, or both [2]. The geometric features capture changes in the location of specific salient facial points caused by activity of facial muscles (*e.g.*, the displacement of the facial points between expressive and expressionless faces [4]). On the other hand, the appearance-based features capture transient differences in the facial appearance such as wrinkles, bulges and furrows. While the former are more robust to illumination and pose changes, not all AUs can be detected solely from the geometric features [5]. For example, the activation of AU6 wrinkles the skin around the outer corners of the eyes and raises the cheeks, which makes it difficult, if not impossible, to detect this AU from facial landmarks only. On the other hand, appearance-based features are typically high-dimensional and contain subject-specific information, both of which can adversely affect the classification/detection performance. Therefore, using both geometric and appearance features might be the best choice, letting the model to choose the most relevant features for detection of target AUs. Thus, our goal is to achieve an effective fusion of these two types of features while still keeping the model computationally tractable.

AUs rarely appear in isolation (more than 7,000 AU combinations have been observed in everyday life [6]). For this reason, the AU detection can be improved at the model level by exploiting the 'semantics' of AUs, in terms of their co-occurrences. These co-occurrences are usually driven by

Manuscript received December 18, 2015; revised July 24, 2016 and September 26, 2016; accepted September 27, 2016. Date of publication October 5, 2016; date of current version October 18, 2016. This work was supported by the European Community Horizon 2020 [H2020/2014-2020] under Grant 645094 (SEWA). The work of S. Eleftheriadis was supported by the European Community 7th Framework Programme [FP7/2007-2013] under Grant 611153 (TERESA). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Amit K. Roy Chowdhury.

S. Eleftheriadis and O. Rudovic are with the Department of Computing, Imperial College London, London, SW7 2AZ, U.K. (e-mail: stefanos@imperial.ac.uk; o.rudovic@imperial.ac.uk).

M. Pantic is with the Department of Computing, Imperial College London, London, SW7 2AZ, U.K., and also with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7522 NB Enschede, The Netherlands (e-mail: m.pantic@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2615288

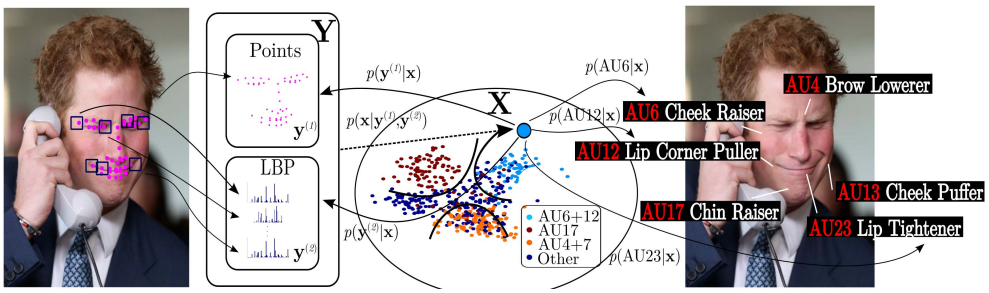


Fig. 1. The proposed MC-LVM. The geometrical and appearance input features, $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$, are first projected onto the shared manifold $\boldsymbol{X}$. The fusion is attained via GP conditionals, $p(\boldsymbol{y}^{(1)}|\boldsymbol{x})$ and $p(\boldsymbol{y}^{(2)}|\boldsymbol{x})$, that generate the inputs. Classification is performed on the manifold via simultaneously learned logistic functions $p(z^{(c)}|\boldsymbol{x})$ for multiple AU detection. The subspace is regularized using constraints imposed on both latent positions and output classifiers, encoding local and global dependencies among the AUs.

the context in which the target facial behavior occurs (e.g., pain or joy). Encoding this type of information during the joint AU prediction helps to reduce the space of possible AU combinations in target data, resulting in simpler and more effective models for the joint prediction. Also, the co-occurring AUs can be non-additive, in the case of which one AU masks another, or a new and distinct set of appearances is created [3]. For instance, AU4 (brow lowerer) appears differently depending on whether it occurs alone or in combination with AU1 (inner brow raise). When AU4 occurs alone, the brows are drawn together and lowered. In AU1+4, the brows are drawn together but are raised due to the action of AU1. This, in turn, significantly affects the appearance features of the target AUs. Moreover, some AUs are often activated together due to the latent variables such as emotions (*e.g.*, AU12 and AU6 in the case of joy).

Despite all this, most of the existing approaches to AU detection model each AU independently, using either a single type of facial features [7], [8], or combining multiple features by means of naive approach (i.e., simple feature concatenation) [4], [9] or multiple-kernel learning (MKL) [10]. Furthermore, some methods treat different combinations of AUs as new independent classes [11]; however, this is impractical given the number of possible AU combinations. On the other hand, methods that do attempt to model the AU co-occurrences (*e.g.*, [12]–[14]) fail to exploit different types of facial features in their models. To our knowledge, the only methods that attempt both are [15]–[17]. However, these methods either suffer from the curse of dimensionality as they perform feature fusion by concatenation of geometric- and appearance-based features using parametric models [15], [16], or cannot model more than a few AUs jointly due to the computational burden of their (non-parametric) inference methods [17].

To this end, we propose a Multi-conditional Latent Variable Model (MC-LVM) that performs simultaneously the fusion of different facial features and joint detection of AUs. Instead of performing the AU detection in the original feature space, as done in existing works [15]–[17], the MC-LVM attains the feature fusion via a low-dimensional subspace shared across feature sets. This subspace is learned by employing the framework of shared Gaussian processes (GPs) [18]. Here, the learning is constrained by two types of newly introduced constraints. *Topological* constraints encode local dependencies (from image pairs) among multiple AUs by means of string kernels [19]. *Relational* constraints, enforce the co-occurrences of the model predictions to match those of the target labels. The learning of the subspace is performed jointly with the AU detectors. The latter are modeled via multiple logistic regressors which operate on the shared subspace of the fused features. Note that, in contrast to existing multi-output subspace learning methods (*e.g.*, [20], [21]), the MC-LVM learns a subspace for multiple AU detection that combines both the generative and discriminative properties of probabilistic models, while simultaneously modeling the AU correlations at both feature level (via the proposed fusion approach) and model level (via the introduced regularizers). Due to its multi-conditional likelihood function, the proposed model is less susceptible to overfitting compared to purely discriminative models. Its generative part acts as an efficient regularizer during parameter learning. The proposed multi-conditional learning is motivated by the fact that discriminative learning usually yields better results when provided with sufficient training data, as it does not expend its modeling power on the marginal distribution of the features, as done in its generative counterparts. On the other hand, generative models, if specified well, can generalize better with fewer training data [22]. Thus, leveraging the advantages of the two approaches during the model learning process is expected to lead to better generalization performance. To further improve the robustness and efficiency of the parameter estimation, a Bayesian learning of the data subspace is facilitated through Monte Carlo sampling, and an Expectation-Maximization (EM)-like learning approach. During inference, the simultaneous detection of multiple AUs is performed by applying the learned back-mappings from inputs to the shared subspace, where the detection of target AUs is performed consequently. The outline of the proposed approach is illustrated in Fig. 1.

To summarize, the contributions of the proposed work are:

- To the best of our knowledge, this is the first approach for multiple AU recognition that jointly performs facial feature fusion and AU detection simultaneously, via manifold learning. The proposed MC-LVM is derived in a fully Bayesian multi-conditional formulation, and combines the properties of *both* generative and discriminative model

by merging the framework of shared GPs with logistic classifiers.
- We introduce novel *topological* and *relational* constraints that successfully encode the AU dependencies at both feature and model level into the proposed manifold learning for joint AU detection. We show that such constraints play an important role in increasing the discriminative power of the learned manifold, resulting in improved (average) detection performance.
- We demonstrate on three publicly-available datasets that the proposed approach outperforms the state-of-the-art methods for joint AU detection, and several recently proposed methods for feature fusion and multi-label classification.

Note that a preliminary version of this work appeared in [23]. Herein, we extensively evaluate the model's performance under various settings, in order to extend and conclude the analysis performed in [23]. Specifically, in our experiments are now included: 1) A thorough assessment of the contribution of the weighted multi-conditional formulation to the detection of each AU. 2) An evaluation of the generalization ability of the proposed model on two cross-dataset scenarios. 3) Additional experimental results based on extra evaluation metrics for our comparisons to existing state-of-the-art.

The remainder of the paper is organized as follows. Sec. II gives an overview of existing work on AU detection and related models for the target task. In Sec. III, we introduce the proposed MC-LVM. Sec. IV shows the results of the experimental evaluation, and Sec. V concludes the paper.

II. Related Work

A. Multiple Facial AU Detection

The majority of the existing works attempt to recognize AUs or certain AU combinations independently [4], [5], [7]–[9], [11], [24], [25]. A common limitation of these methods is that they construct independent AU classifiers that ignore the relations among the AUs. Based on how the AU-specific classifiers are designed, they can be divided into two main categories: (a) *static* modeling approaches, where each frame is evaluated independently [4], [7]–[9], [11], and (b) *temporal* modeling approaches, where temporal dynamics are explored within a video sequence [5], [24], [26]. Representatives of the first group commonly apply independent classifiers, *e.g.*, support vector machine (SVM) [4], [9], and Adaboost [7] on the collected features, or use the notion of domain adaptation to develop personalized AU-classifiers [8]. Alternatively, in [11] sparse representations are employed to create a dictionary of facial images with certain AU combinations. In the second group, the majority of the works are based on variants of dynamic Bayesian networks (DBN). Reference [5] combines SVM and hidden Markov models (HMM) to encode the AUs and their temporal activations, while the authors in [24] use a combination of GentleBoost and HMM for the target task. More recently, the authors in [26] account for the ordinal information in the framework of conditional random field (CRF), to model the relations between the temporal segments of each AU. Regardless of the modeling technique, none of the above methods takes into account the relations among the AUs.

To the best of our knowledge, there are only few works that perform joint detection of AUs [12], [13], [15]–[17], [27]. Reference [12] proposed a generative framework based on DBNs to model the semantics of different AUs. Due to the Markov assumptions while learning the network of the co-occurred AUs, this model can handle only local dependencies between pairs of AUs. The authors in [27] propose a generative latent tree algorithm for AU intensity estimation. The dependencies among observed features and multiple AUs are modeled via latent variables. Nevertheless, [12] and [27] lack the classification power of the discriminative models. On the other hand, the models in [13] and [15]–[17] are defined in a fully discriminative framework. Specifically, [16] first learns the logistic classifiers for multiple AUs using the notion of multi-task feature learning, and then uses a pre-trained BN to refine the predictions. Note that this model fails to account for AU dependencies at the feature level, which can result in loss of information, *e.g.*, in case of non-additive AUs. Reference [14] tries to learn independent logistic classifiers by first selecting a sparse subset of facial patches which are more relevant to each AU. Yet, the fusion task is not addressed, while the AU-dependencies are regarded only between predefined pairs. Reference [15] employed the restricted Boltzmann machine (RBM) to overcome the pair-wise AU modeling limitation of the DBN [12]. The authors proposed a parametric model, in which *discrete* latent variables account for correlations among discrete outputs that are directly connected to the image features. Since the latent variables are not connected to the feature space, they cannot model correlations between the inputs, hence, *concatenation* of the input features is used for the fusion task. Reference [17] combines multi-task learning with MKL to jointly learn different AU classifiers. The authors introduce l_p-norm regularization to the MKL, in order to fuse multiple types of features with different kernels, and account for dependencies among different tasks (*i.e.*, AUs). However, this non-parametric method can deal only with small subsets of AUs (typically less than 4) in its output. Reference [13] proposed a probabilistic framework, based on Bayesian compressed sensing (BCS), to encode the co-occurrence structure and the (group) sparsity patterns of the AUs to the compressed signal (latent variables). The relations between the original data and the latent variables are modeled via linear regression, where the inputs are the appearance based features. Hence, this work cannot deal with fusion of different input features.

The proposed approach advances the existing work in many aspects. The fusion of the facial features is performed in a continuous (low-dimensional) subspace, allowing the model to capture dependencies among multiple AUs at both feature and model level during learning. Contrary to the methods mentioned above, which are purely generative or discriminative, the proposed MC-LVM takes the best of both approaches and successfully combines them in its multi-conditional likelihood function. Note that the the proposed MC-LVM is closely related to the MKL model in [17], which performs the feature fusion implicitly via the kernel-induced space, while MC-LVM does it explicitly via the fixed point estimate of the

shared low-dimensional latent projections. Yet, the complexity of the model in [17] increases quadratically with the number of AUs in the output, while it increases only linearly in case of MC-LVM. Consequently, MC-LVM can efficiently model relations among a relatively large number of outputs, without the requirement to *a priori* define groups of AUs as done in [14] and [17].

B. Multi-Modal Fusion

The analysis of multi-modal streams of data has attracted significant research attention in the fields of computer vision and especially the facial behavior analysis. Reference [28] provides an extensive overview on how vocal, gestural and facial features extracted from both audio and visual modalities can be used to identify particular human behaviors. As we have already seen, the most evident way towards feature fusion is to concatenate the individual modalities and apply a single classifier for the target task [4], [9]. An orthogonal approach is to first train individual classifiers per modality and then fuse the predictions, *e.g.*, by feeding their outputs in another classifier as in [29]. Alternatively, fusion can be performed via employing the framework of MKL, which aims to integrate the information from different features by learning a weighted combination of respective kernels. A detailed survey of the MKL-related methods can be found in [30]. Another possible direction toward feature fusion is to exploit the notion of joint sparse representations or learn multi-modal dictionaries based on joint sparsity constraints. Based on these techniques, the authors in [31] and [32] managed to fuse the information from various biometrics in order to perform more accurate face verification. A similar approach is to perform joint dimensionality reduction and project the multiple features on a common subspace. For instance, in [33] and [34] facial images from various channels (*e.g.*, infrared images and forensic sketches) are commonly projected to the space obtained by PCA, before applying a classifier for face recognition. Likewise, the authors in [35] employed the canonical correlation analysis (CCA) in order to fuse the information from fMRI, sMRI and EEG data, for detecting patients diagnosed with schizophrenia.

In the current work we follow the approach of joint dimensionality reduction in order to fuse the information from the geometric and appearance features. In contrast to the methods described above, MC-LVM employs the framework of shared GP latent variable models (S-GPLVM) [18] to unravel a shared non-linear manifold that generates the input features. This results in a more natural fusion, since the latent representations are learned in a way to to generate the multiple modalities. The generative process of MC-LVM is utilized via a non-parametric probabilistic mapping from the latent space to the observed features. This property constitutes the proposed approach less prone to overfitting. Finally, MC-LVM, as an inherent kernel method, can effectively deal with input features of higher dimensionality and more complex structure.

C. Multi-Label Classification

The proposed MC-LVM is related to existing works on multi-label classification that attempt to learn robust classifiers by exploiting efficiently the label dependencies. For an extensive overview, the reader is referred to [36] and [37]. For instance, [38] extended the k-nearest neighbor (kNN) to the multi-label scenario by using the number of neighboring instances belonging to each possible class, as prior information to determine the label set for an unseen instance. Reference [39] derived the back-propagation algorithm of the neural networks for the multi-label classification. Reference [40] proposed an approximate learning approach in order to extended the work of structured SVM [41] to multi-label classification. The latter is also highly related to multi-task learning techniques. The latter rely on the introduction of an inductive bias on the joint space of all tasks (*e.g.*, AUs) that reflects our prior beliefs regarding the related structure. A popular approach is to jointly learn the tasks under a regularization framework [42]. The regularization operates on the parameter space and penalizes distances between the different tasks, which results in uncovering a common set of parameters across the tasks. Hence, it allows to capture the similarities among the outputs through parameter sharing. Based on this idea, [21] introduced a manifold regularization approach to the multi-task learning. The key assumption is that the task parameters lie on a low dimensional manifold, and thus, they cannot vary arbitrarily. Instead of explicitly learning the manifold, the authors model the projection functions in a parametric formulation, and alternate between solving for the task parameters and minimizing their distances in the projected manifold. Similarly, [20] defines a latent variable model, which generates the task specific parameters in a probabilistic fashion. Due to its probabilistic formulation, several priors can be imposed on the latent variables to induce a desired structure to the task specific manifold.

The above methods rely on implicit assumptions that all tasks are related to each other. Contrary to this belief, [43] aims to uncover a structured pattern among the tasks, and combine them into different groups. Each task parameters are assumed to be a sparse, linear combination of underlying latent basic tasks. The overlap in the sparsity patterns of any two tasks controls the amount of sharing between them. In a similar fashion, [44] introduced the use of multi-output GP, for modeling task dependent regressors (latent functions) via GP priors. The output of each task is a weighted combination of a number of shared latent functions, which enables the collaboration among the tasks, plus an individual task-specific latent function. In order to deal efficiently with the problem of large number of output tasks and input data points, the authors derived a formulation based on variational inference. Following a different approach, [45] used the notion of spectral graph regularization to jointly learn clusters of closely related tasks. Relationships between the tasks are defined in terms of the graph Laplacian, which favors similar tasks to be close in the parameter space. The authors proposed an alternating optimization algorithm based on proximity operators, in order to jointly learn the tasks and the graph. While applicable to the task of multiple AU detection, these methods do not perform simultaneous feature fusion and multi-label classification. By contrast, the proposed MC-LVM can be seen as a multi-task learning approach, where the relations of different tasks (*i.e.*, AUs) are learned directly in the shared subspace, by

implicitly relating them through their feature and label dependencies. The latter are encoded by the local and global priors proposed in our model.

More recent works in the GP and multi-label classification context [46], [47] try to combine multi-task learning and feature fusion via subspace learning. Reference [46] jointly optimizes latent variables in order to reconstruct the input data, and account for multiple tasks in the output. A downside of this method is that latent space learning is done by the maximum likelihood (ML) strategy, *i.e.*, the latent space is directly optimized during learning. In the case of large amount of data, this can easily lead to overfitting [48]. To ameliorate this, [47] proposed a fully Bayesian framework, based on variational inference, to integrate out the latent space.

In contrast to these methods, MC-LVM employs multi-conditional learning strategies to re-weight the generative and discriminative conditionals, in order to unravel a suitable subspace for joint feature fusion and multi-label classification. In our Bayesian approach, the latent space is approximated via an efficient Monte Carlo sampling, where the conditional models determine the importance of each sample. Moreover, the inference step is efficiently performed via the learned projection mappings to the manifold. This overcomes the requirement of [47] to learn another approximation to the posterior of the test inputs. Finally, note that none of these approaches have been evaluated in the task of multiple AU detection.

III. Multi-Conditional Latent Variable Model

A. Notation and Preliminaries

Let us denote the training set as $\mathcal{D} = \{\boldsymbol{Y}, \boldsymbol{Z}\}$, which is comprised of V observed input channels $\boldsymbol{Y} = \{\boldsymbol{Y}^{(v)}\}_{v=1}^{V}$, and the associated output labels $\boldsymbol{Z}$. Each observed channel is comprised of N i.i.d. multivariate samples $\boldsymbol{Y}^{(v)} = \{\boldsymbol{y}_i^{(v)}\}_{i=1}^{N}$, where $\boldsymbol{y}_i^{(v)} \in \mathbb{R}^{D_v}$ denote corresponding facial features. Furthermore, $\boldsymbol{Z} = \{z_i\}_{i=1}^{N}$ denote multiple binary labels, with $z_i \in \{-1, +1\}^C$ encoding C (co-occurring) outputs. Let us further assume the existence of a latent space $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathcal{R}^q$, $q \ll D_v$, which is a low-dimensional representation of the original observations $\boldsymbol{Y}$. This implies that there exists a set of latent functions $f^{(v)}$, that can generate $\boldsymbol{y}_i^{(v)}$ from $\boldsymbol{x}_i$, *i.e.*, $\boldsymbol{y}_i^{(v)} = f^{(v)}(\boldsymbol{x}_i) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_v^2 \boldsymbol{I})$ is additive Gaussian noise. In the proposed approach we model these functions using the framework of GPs [19]. For notation simplicity, we set the number of input spaces to $V = 2$, as generalization to more than two input spaces is straightforward. The model outline is depicted in Fig. 2.

B. Model Definition

Our goal is to learn a model that simultaneously combines different inputs and detects activations of multiple outputs. To this end, we are interested in finding the latent representations $\boldsymbol{x}$, that jointly generate $\boldsymbol{y}$ and $\boldsymbol{z}$. In a Bayesian approach, this requires the computation of the joint marginal likelihood:

$$p(\boldsymbol{y}, z) = \int p(\boldsymbol{y}^{(1)}|\boldsymbol{x})p(\boldsymbol{y}^{(2)}|\boldsymbol{x})p(z|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}, \qquad (1)$$

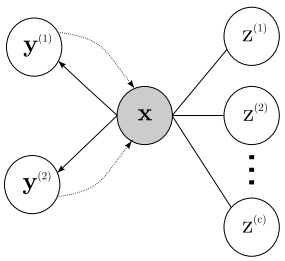


Fig. 2. Graphical representation of the proposed MC-LVM. The definition of the conditionals is given in Sec. III-C.

where we exploited the property of conditional independence, *i.e.*, $\{\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, z\}$ are independent given the latent variable $\boldsymbol{x}$. Note that in order to compute the above integral, we need to marginalize out $\boldsymbol{x}$. However, for the non-linear conditional models, which we detail in Section III-C, the integral in Eq. (1) is intractable. To overcome this, we numerically approximate the marginal likelihood using Monte Carlo sampling [49]

$$p(\boldsymbol{y}, z) \approx \frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{y}^{(1)}|\boldsymbol{x}_s)p(\boldsymbol{y}^{(2)}|\boldsymbol{x}_s)p(z|\boldsymbol{x}_s). \qquad (2)$$

The samples $\boldsymbol{x}_s$, $s = 1, \ldots, S$ are drawn from $p(\boldsymbol{x})$, which is defined in Sec. III-C. Using the Bayes' rule, we can derive the posterior over the latent variable

$$p(\boldsymbol{x}|\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, z) = \frac{p(z|\boldsymbol{x})p(\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}|\boldsymbol{x})p(\boldsymbol{x})}{\frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}|\boldsymbol{x}_s)p(z|\boldsymbol{x}_s)}. \qquad (3)$$

We then calculate the above probability for all pairs of training data i and Monte Carlo latent samples s, to obtain the membership probabilities $p(s, i) = p(\boldsymbol{x}_s|\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_i^{(2)}, z_i)$. Hence, $p(s, i)$ denotes the posterior probability of acquiring the sample $\boldsymbol{x}_s$, having observed the inputs $\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_i^{(2)}$ and outputs z_i. This gives rise to the expectation of the latent points under the sampling distribution:

$$\boldsymbol{x}_i = E\{\boldsymbol{x}|\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_i^{(2)}, z_i\} = \sum_{s=1}^{S} p(s, i)\boldsymbol{x}_s, \qquad (4)$$

which allows us to obtain the point estimates of the shared latent positions without explicitly optimizing them for each training pair. In this way, not only we end up with a probabilistic estimation of the latent space, but we also considerably reduce the number of model parameters.

C. Conditional Models

From Eq. (1), we see that the marginal likelihood of the desired model is composed of the conditional probabilities $p(\boldsymbol{y}^{(v)}|\boldsymbol{x})$ and $p(z|\boldsymbol{x})$, while it also depends on the sampling distribution $p(\boldsymbol{x})$. Hence, the correct choice of these distributions affects critically the representational capacity of the shared subspace, and thus, the model's performance. Effectively, this requires the learning of the conditional models that provide: (i) *generative* mappings from the latent space to the inputs ($\boldsymbol{x} \rightarrow \boldsymbol{y}^{(v)}, v = 1, 2$); (ii) *projection* mappings from the inputs to latent space ($\boldsymbol{y}^{(v)} \rightarrow \boldsymbol{x}$); (iii) *discriminative* mappings

from latent space to multiple binary outputs ($\boldsymbol{x} \rightarrow \boldsymbol{z}$), as depicted in Fig. 2.

1) Generative Mappings: Different probabilistic models such as Gaussian models [50] or naive Bayes models [51] can be employed to recover the generative mappings. Yet, parametric models are limited in their ability to recover non-linear mappings from the latent space to high-dimensional input features. To this end, we place GP priors on the functions that generate the observed features. This gives rise to the likelihood:

$$p(\boldsymbol{Y}^{(\upsilon)}|\boldsymbol{X}, \boldsymbol{\theta}^{(\upsilon)}) = \frac{1}{\sqrt{(2\pi)^{ND_\upsilon}|\boldsymbol{K}_{\boldsymbol{Y}}^{(\upsilon)} + \sigma_\upsilon^2 \boldsymbol{I}|^{D_\upsilon}}} \times \exp\left[-\frac{1}{2}\mathrm{tr}\left((\boldsymbol{K}_{\boldsymbol{Y}}^{(\upsilon)} + \sigma_\upsilon^2 \boldsymbol{I})^{-1}\boldsymbol{Y}^{(\upsilon)}\boldsymbol{Y}^{(\upsilon)T}\right)\right], \quad (5)$$

where $\boldsymbol{K}_{\boldsymbol{Y}}^{(\upsilon)}$ is a $N \times N$ kernel matrix, obtained by applying the covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ to the elements of $\boldsymbol{X}$, and it is shared across the dimensions of $\boldsymbol{Y}^{(\upsilon)}$. The covariance function is usually chosen as the sum of the radial basis function (RBF) kernel, bias and noise terms

$$k^{(\upsilon)}(\boldsymbol{x}, \boldsymbol{x}') = \theta_1^{(\upsilon)} \exp(-\frac{\theta_2^{(\upsilon)}}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2) + \theta_3^{(\upsilon)} + \frac{\delta_{\boldsymbol{x},\boldsymbol{x}'}}{\theta_4^{(\upsilon)}}, \quad (6)$$

where $\delta_{\boldsymbol{x},\boldsymbol{x}'}$ is the Kronecker delta function, and $\boldsymbol{\theta}^{(\upsilon)} = \{\theta_1^{(\upsilon)}, \theta_2^{(\upsilon)}, \theta_3^{(\upsilon)}, \theta_4^{(\upsilon)}\}$ are the kernel hyperparameters. The parameter learning in GPs is performed by gradient-based minimization of $-logp(\boldsymbol{Y}^{(\upsilon)}|\boldsymbol{X}, \boldsymbol{\theta}^{(\upsilon)})$ w.r.t. $\boldsymbol{\theta}^{(\upsilon)}$ [19]. Then, the predictive probability of the GP for a new $\boldsymbol{x}_*$ is given by

$$p(\boldsymbol{y}_*^{(\upsilon)}|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{Y}^{(\upsilon)}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{y}_*^{(\upsilon)}}, \sigma^2_{\boldsymbol{y}_*^{(\upsilon)}}), \quad (7)$$

with $\boldsymbol{\mu}_{\boldsymbol{y}_*^{(\upsilon)}}$ and $\sigma^2_{\boldsymbol{y}_*^{(\upsilon)}}$ as:

$$\boldsymbol{\mu}_{\boldsymbol{y}_*^{(\upsilon)}} = \boldsymbol{k}_*^{(\upsilon)T}(\boldsymbol{K}_{\boldsymbol{Y}}^{(\upsilon)} + \sigma_\upsilon^2 \boldsymbol{I})^{-1}\boldsymbol{Y}^{(\upsilon)} \quad (8)$$

$$\sigma^2_{\boldsymbol{y}_*^{(\upsilon)}} = k_{**}^{(\upsilon)} - \boldsymbol{k}_*^{(\upsilon)T}(\boldsymbol{K}_{\boldsymbol{Y}}^{(\upsilon)} + \sigma_\upsilon^2 \boldsymbol{I})^{-1}\boldsymbol{k}_*^{(\upsilon)} + \sigma_\upsilon^2. \quad (9)$$

The kernel values $\boldsymbol{k}_*^{(\upsilon)}$ and $k_{**}^{(\upsilon)}$ are computed by applying Eq. (6) to the pairs $(\boldsymbol{X}, \boldsymbol{x}_*)$ and $(\boldsymbol{x}_*, \boldsymbol{x}_*)$, respectively, and σ_υ^2 is the noise of the process. Hence, the conditional model $p(\boldsymbol{y}^{(\upsilon)}|\boldsymbol{x}), \upsilon = 1, 2$, in Eq. (3) is now fully defined by the Gaussian distribution in Eq. (7), where the latent sample $\boldsymbol{x}_s$ acts as the new latent position $\boldsymbol{x}_*$.

2) Projection Mappings and Sampling: To model the sampling distribution $p(\boldsymbol{x})$, the simplest choice is to assume a spherical Gaussian prior over the latent points $\boldsymbol{x}$. However, such an uninformative prior would give rise to latent representations that cannot effectively exploit the structure of input data. To this end, we define a sampling distribution that constraints the samples $\boldsymbol{x}_s$ by conditioning them on the inputs, *i.e.*, $\tilde{p}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)})$. This is motivated by the notion of back-constraints in GP latent variable model (GPLVM) [52], where this type of conditional distribution is used to learn the mappings from the input to the latent space. We learn the conditional model for $\tilde{p}(\boldsymbol{x})$ using GPs, as done for the generative mappings. The use of GPs in the projection mappings, apart from modeling the sampling distribution, also allows us to easily combine multiple features within its kernel matrix as $\boldsymbol{K}_X = \boldsymbol{K}_X^{(1)} + \boldsymbol{K}_X^{(2)}$, corresponding to the sum of the kernel functions defined on $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$, respectively. Hence, the resulting kernel is responsible for effectively performing the non-linear fusion of the input features into a single latent point. It can be regarded as an automatic MKL approach with non-parametric GP regression functions. Finally, the resulting conditional model $p(\boldsymbol{x}_*|\boldsymbol{y}_*^{(1)}, \boldsymbol{y}_*^{(2)})$ has the form of Eq. (7) (with the relations between $\boldsymbol{y}^{(\upsilon)}$ and $\boldsymbol{x}$ being reverted), and since it is a low-dimensional Gaussian distribution, sampling from it can be performed efficiently.

3) Discriminative Mappings: Since we are interested in binary detection of activations of multiple AUs, we use the conditional models based on the logistic regression [19] to model $p(\boldsymbol{z}|\boldsymbol{x})$. By assuming conditional independence given the latent positions $\boldsymbol{x}$, we can factorize this conditional as:

$$p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{W}) = p(z^{(1)}|\boldsymbol{x}, \boldsymbol{w}_1) \ldots p(z^{(C)}|\boldsymbol{x}, \boldsymbol{w}_C), \quad (10)$$

$$p(z^{(c)}|\boldsymbol{x}, \boldsymbol{w}_c) = (1 + e^{-\boldsymbol{x}^T \boldsymbol{w}_c z^{(c)}})^{-1}, \quad c = 1, \ldots, C, \quad (11)$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_C] \in \mathcal{R}^{q \times C}$ contains the weight vectors of the individual functions. During inference, if $p(z_*^{(c)}|\boldsymbol{x}_*) > 0.5$, the c-th output is active, *i.e.*, $z_*^{(c)} = 1$.

D. Output Constraints

Due to the potentially large number of outputs, the topology of the latent space needs to be constrained to avoid the model focusing on unimportant variation in the data (*e.g.*, modeling relations between rarely co-occurring outputs). Furthermore, we need to encourage the model to produce similar predictions for outputs that are more likely to co-occur (*e.g.*, AU6+12), and competing predictions for those that rarely co-occur (*e.g.*, AU12 and AU17). We describe below how we construct target constraints based on the output relations, and how these are incorporated into the MC-LVM framework.[1]

1) Topological Constraints: Herein, we define the constraints that encode co-occurrences of the output labels using the notion of graph regularization [53]. We construct an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$ is the node set, with node V_i corresponding to latent positions $\boldsymbol{x}_i$, and $\mathcal{E} = \{(V_i, V_j)_{i,j=1\ldots N} | i \neq j, \boldsymbol{x}_i$ and $\boldsymbol{x}_j$ have co-activated outputs$\}$ is the edge set. By pairing each node with the latent variables, we obtain a Gaussian Markov random field [54] w.r.t. graph $\mathcal{G}$. Next, we need to associate each edge in the graph with a weight. For this purpose we encode the relations between the data into an $N \times N$ weight matrix $\boldsymbol{S}$. The latter is defined in a supervised fashion by measuring the similarity between the output label vectors using the notion of string kernels [19] as:

$$\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{l \in \mathcal{A}} \boldsymbol{z}_{l,x}^T \boldsymbol{z}_{l,x'}, \quad (12)$$

where $\mathcal{A}$ is the set of all possible 2^C combination of the output labels and l is the set of possible sub-labels of tuples, triples,

[1]For the mathematical analysis of this subsection, the negative class in the output labels z will be denoted with 0 instead of the used -1.

etc. $\boldsymbol{z}_{l,x}$ denotes the specific sub-label of $\boldsymbol{x}$ and holds the currently active 'sub-string' l of the actual labels. Hence, $\boldsymbol{S}_{ij}$ contains the number of co-activated outputs in all sub-labels between two instances. Note that contrary to [14], we measure the similarity of the outputs based on all possible groups of co-occurring AUs, and not only on pairs of AUs. The graph Laplacian matrix is then defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, where $\boldsymbol{D}$ is a diagonal matrix with $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{S}_{ij}$. Finally, using Eq. (4), we arrive at the Laplacian regularization term

$$C = \mathrm{tr}(\boldsymbol{X}^T \boldsymbol{L} \boldsymbol{X}) = \sum_{i,j}^{N} \sum_{s=1}^{S} \sum_{t=1}^{S} L_{ij}\, p(s,i)\, p(t,j) \boldsymbol{x}_s^T \boldsymbol{x}_t. \quad (13)$$

Eq. (13) incurs higher penalty if latent projections of co-occurring AUs are distant in the manifold. Thus, projections with strongly related AUs are placed close to each other.

2) Global Relational Constraints: In order for the MC-LVM to fully benefit from the above topological constraint, it is important to ensure that the model produces similar predictions for frequently co-occurring AUs. Therefore, we introduce the global relational constraints as:

$$R = \| \boldsymbol{P}_z^T \boldsymbol{P}_z - \boldsymbol{Z}^T \boldsymbol{Z} \|_F^2, \quad (14)$$

where $\boldsymbol{P}_z = [p(z_1|\boldsymbol{x}_1), \ldots, p(z_N|\boldsymbol{x}_N)]^T$ are the predictions from Eq. (11) for each $\boldsymbol{x}_i$, and $\boldsymbol{Z}$ is the true label set. Thus, Eq. (14), incurs a high penalty if correlated outputs have dissimilar predictions. In this way, the co-occurrence matrix of the predictions is forced to be similar to that of the true labels, and hence, the discriminative power of the output detectors is increased.

E. Learning and Inference

The objective function of our model is the sum of the complete data log-likelihood of the (weighted) joint distribution in Eq. (2) penalized by the constraints in Eq. (13,14)

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \log \sum_{s=1}^{S} \underbrace{p(\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}|\boldsymbol{x}_s)}_{p_{g,i}}{}^{1-\alpha} \underbrace{p(\boldsymbol{z}|\boldsymbol{x}_s)}_{p_{d,i}}{}^{\alpha} - \lambda_C C - \lambda_R R, \quad (15)$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(v)}, \boldsymbol{W}\}$. Note that in contrast to the standard ML optimization, we set the parameter $\alpha \in [0, 1]$ to find an optimal balance between the generative ($p_{g,i}$) and discriminative ($p_{d,i}$) components of our MC-LVM. The generative component has the key role in unraveling the latent space of the fused features, while the discriminative component regularizes the manifold by using the labels' structure information. Large α values give rise to models that depend more on the labels to define the decision boundaries for the detection, while for small α the model expends more effort on capturing the variations in the features (*e.g.*, due to various sources of noise in data such as head-pose variation in spontaneous data). By finding optimal α via a cross-validation procedure, as explained in Sec. IV-C, we allow the model to find a trade-off between the discriminative and generative part.

Another key difference to the ML approach, is that the Bayesian optimization requires the computation of the posterior of the latent space. The latter depends on the parameters $\boldsymbol{\Theta}$, and thus, direct optimizing of the objective in Eq. (15) w.r.t. $\boldsymbol{\Theta}$ is not possible. Hence, we propose an EM-based approach for parameter learning. In the E-step, we find the expectation of the complete-data log-likelihood in Eq. (15) under the posterior in Eq. (3), which is given by

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(old)}) = \sum_{i=1}^{N} \sum_{s=1}^{S} p(s,i) \log \left(p_{g,i}^{1-\alpha} p_{d,i}^{\alpha} \right), \quad (16)$$

where the membership probabilities, $p(s,i)$, are computed with $\boldsymbol{\Theta}^{(old)}$. In the M-step, we find $\boldsymbol{\Theta}^{(new)}$ by optimizing

$$\boldsymbol{\Theta}^{(new)} = \arg\max_{\boldsymbol{\Theta}} \quad Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(\mathrm{old})}) - \lambda_C C - \lambda_R R, \quad (17)$$

w.r.t. $\boldsymbol{\Theta}$ using the conjugate gradient method [19].

The full training of the model is split into two stages, where in each stage we compute $p(\boldsymbol{x}|\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)})$ and $p(\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, z|\boldsymbol{x})$ in an alternating fashion. Specifically, we first initialize the latent coordinates $\boldsymbol{X}$, using a dimensionality reduction method, *e.g.*, PCA [49], on the concatenation of the two feature sets. Then, we learn the sampling distribution $p(\boldsymbol{x}|\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)})$ by training a GP on the projection mappings, as explained in Sec.III-C, and collect S samples from corresponding GP posterior. During the second stage, we employ the EM algorithm described above to learn the parameters $\boldsymbol{\Theta}$. Note that the constraints C and R implicitly depend on the posterior, which is a function of the current estimate of $\boldsymbol{\Theta}$, hence, we need to compute their derivatives w.r.t to $\boldsymbol{\Theta}$. The penalized log-likelihood can be optimized jointly [50] or separately [55] without violating the EM-optimization scheme, since the updates from the penalty terms do not affect the computation of the expectation. After the M-step we refine our original estimate of the latent space $\boldsymbol{X}$, using Eq. (4). We iterate between stage 1 and 2 until convergence of the objective function in Eq. (17).

1) Inference: Inference in the proposed MC-LVM is straightforward. The test data $\boldsymbol{y}_*^{(1)}$, $\boldsymbol{y}_*^{(2)}$, are first projected onto the manifold using Eq. (7). In the second step, the activation of each output is detected by applying the classifiers from Eq. (11) to the obtained latent position. The learning and inference procedure described above is summarized in Alg. 1.

2) Theoretical Analysis: The optimization scheme described earlier in this section does not have theoretical guarantees that it increases the penalized complete log-likelihood after each EM cycle. The reasons behind this are twofold: (i) Eq. (17) cannot be solved analytically, and thus, we need to resort to an iterative procedure based on the conjugate gradient method. Therefore, in each M-step we can only guarantee that a local optimum of the posterior will be recovered. (ii) The expectation of the complete log-likelihood in Eq. (16) is numerically approximated via Monte Carlo sampling, and thus, as in every stochastic optimization problem there is no guarantee that the objective function will strictly increase after each iteration. Hence, it is required to take cautious steps in order not to derive diverge solutions. By carefully initializing both the latent coordinates and the kernel hyper-parameters, and appropriate selection of the number of samples, S, we can effectively learn a latent space with correctly recovered data structure. This is illustrated in Fig. 3, where we can see how the topological constraint

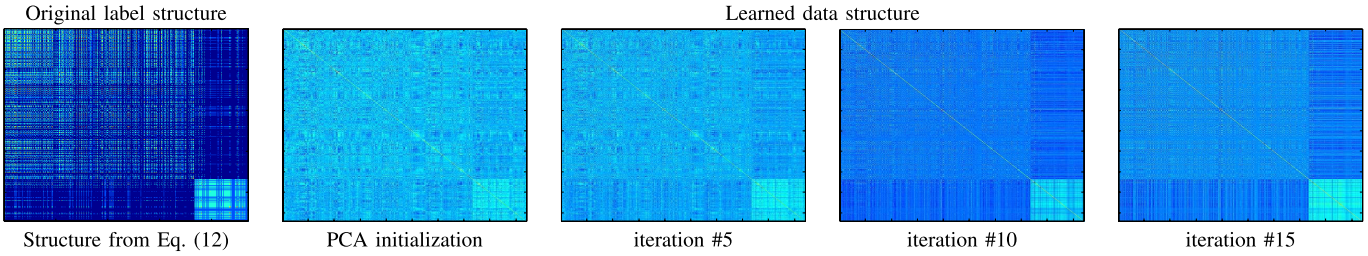


Fig. 3. Evolution of the learned data structure in $\boldsymbol{K}_{\boldsymbol{Y}^{(1)}}$, through the EM-iterations during the optimization on CK+ dataset. The kernels are sorted in order to depict the structure of AU12 (bottom right square) compared to other AU activations (upper right square).

Algorithm 1 MC-LVM: Learning and Inference

Learning
Inputs: $\mathcal{D} = (\boldsymbol{Y}^{(v)}, \boldsymbol{Z}), v = 1, \ldots, V$
Initialize $\boldsymbol{X}$ using PCA on the concatenated $\boldsymbol{Y}^{(v)}$.
repeat
  Stage 1
    Learn $\tilde{p}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)})$ by training the specified GP.
    Draw S samples $\boldsymbol{x}_s$ from the Gaussian distribution $\tilde{p}(\boldsymbol{x})$.
  Stage 2
    E-step: Use the current estimate of the parameters $\boldsymbol{\Theta}^{(old)}$ to compute the membership probabilities in Eq. (3).
    M-step: Update $\boldsymbol{\Theta}$ by maximizing Eq. (17).
  Stage 3
    Update the latent space using Eq. (4).
until convergence of Eq. (17).
Outputs: $\boldsymbol{X}$, $\boldsymbol{\Theta}$

Inference
Inputs: $\boldsymbol{y}_*^{(1)}, \boldsymbol{y}_*^{(2)}$
Step 1: Find the projection $\boldsymbol{x}_*$ to the latent space using Eq. (8).
Step 2: Apply the logistic classifiers from Eq. (11) to the obtained embedding to compute the outputs $\boldsymbol{z}_*$.
Output: $\boldsymbol{z}_*$

imposes the structure of AU12 on the manifold, through the evolution of the iterative EM algorithm. In the initialization step the latent space can roughly model the structure of the positive class (AU12). As the EM iterations progress we see that MC-LVM not only uncovers the structure of AU12 (iteration #5), but it also differentiates it from the structure of the remaining AUs (iteration #15). Additional experimental evaluations regarding the convergence of MC-LVM are given in Sec. IV-C.

3) Complexity: Since MC-LVM is based on the framework of GPs, each iteration during training (within an EM cycle) requires $\mathcal{O}(N^3)$ computations. On the other hand, inference for a new test sample is far more efficient and can be achieved in real-time, since the evaluation of the predictive mean requires $\mathcal{O}(N)$ (predictive variance is not required for classifying a new test point).

F. Relation to Multi-Conditional Models and GPLVM

In the proposed MC-LVM, we employ the GP framework to derive a latent variable model with a joint distribution given by Eq. (1). We then introduce a set of conditional distributions (observed variables given latent positions $p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x})$, and latent positions given the observed data $p(\boldsymbol{x}|\boldsymbol{y})$) to form the multi-conditional objective function. The idea of multi-conditional learning has originally been explored in [50] and [51]. However, these approaches are based on simple parametric conditional models and can deal with single-input single-output scenarios only. The proposed MC-LVM is a generalization of these approaches to multi-input multi-output settings and non-parametric conditionals modeled via GPs.

Modeling of the aforementioned conditionals in MC-LVM resembles that in the GPLVMs [56]. Namely, manifold relevance determination (MRD) [47], multi-task latent GP (MT-LGP) [46] and discriminative shared GP latent variable model (DS-GPLVM) [57], as purely generative methods, try to model the joint likelihood

$$p(\boldsymbol{Y}, \boldsymbol{X}) = p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X}). \tag{18}$$

The learning in these methods consists of maximizing the (marginal) log-likelihood of the joint given above. References [46] and [57] directly optimize the latent variables in a maximum a posterior (MAP) estimation. Reference [47] maximizes a lower bound of the log-marginal likelihood, which is obtained through a variational distribution that approximates the latent space. By contrast, in MC-LVM we model the distribution of both observed inputs and latent variables by employing the predictive posterior of the GP. This results in learning a more robust mapping $\boldsymbol{x} \rightarrow \boldsymbol{y}$, and also allows us to efficiently estimate an instance of the latent space using the Monte Carlo sampling.

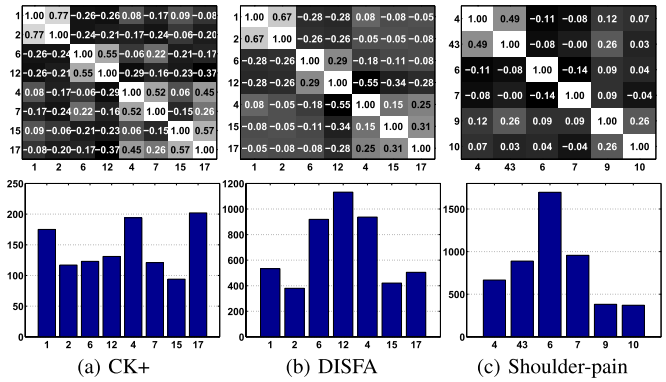
Finally, our proposed sampling distribution is closely related to the notion of 'back-constraints' in the GP literature. Back-constraints were introduced in [52] as a deterministic, parametric mapping that pairs the latent variables of the GPLVM with the observations. This mapping facilitates a fast inference mechanism and enforces structure preservation in the manifold. The same mechanism has been used in [46] and [57]. On the contrary, MC-LVM learns probabilistic mappings via the non-parametric GPs, which can result in latent projections, that are less prone to overfitting.

IV. Experiments

A. Datasets

We evaluate the proposed model on three publicly available datasets: Extended Cohn-Kanade (CK+) [4], UNBC-McMaster Shoulder Pain Expression Archive (Shoulder-pain) [58], and Denver Intensity of Spontaneous Facial Actions (DISFA) [59]. These are benchmark datasets

Fig. 4. Example images with activated AUs from CK+ (top), DISFA (middle) and Shoulder-pain (bottom) datasets.



(a) CK+ (b) DISFA (c) Shoulder-pain

Fig. 5. The global AU relations (in terms of correlation coefficients) (upper row), and the distribution of AU activations within the datasets (lower row).

of posed (CK+), and spontaneous (Shoulder-pain, DISFA) data, containing a large number of FACS coded AUs.

CK+ dataset [4] contains 593 video recordings of 123 subjects displaying posed facial expressions in near frontal views. The image sequences begin from neutral and proceed to the target expression. The last frame (peak frame) is annotated in terms of AU activations (presence/absence). For our experiments, we used the peak frames of all available subjects.

The Shoulder-pain dataset [58] contains video recordings of 25 patients suffering from chronic shoulder pain while performing a range of arm motion tests. Each frame is coded in terms of AU intensity on a six-point ordinal scale.

DISFA dataset [59] contains video recordings of 27 subjects while watching YouTube videos. Again, each frame is coded in terms of the AU intensity on a six-point ordinal scale.

For both DISFA and Shoulder-pain datasets, we treated each AU with intensity larger than zero as active. Sample images from the three datasets, along with examples of AUs present, are shown in Fig. 4. Fig. 5 depicts the AU relations, and the distribution of the AU activations for the data used from each dataset. Note that the co-occurrence patterns and the relations among the AUs differ significantly across all three datasets.

TABLE I

DEFINITIONS OF THE USED AUS FROM CK+, DISFA, AND SHOULDER-PAIN DATASETS

AU	Definition	AU	Definition	AU	Definition
1	Inner brow raiser	7	Lid tightener	15	Lip corner depress.
2	Outer brow raiser	9	Nose wrinkler	17	Chin raiser
4	Brow lowerer	10	Upper lip raiser	43	Eyes closed
6	Cheek raiser	12	Lip corner puller		

B. Experimental Settings

1) Features: In each frame of an input sequence 49 fiducial facial points were extracted using the 2D Active Appearance Model [60]. Based on these points, we registered the images to a reference face (average face for each dataset) using an affine transformation. As input to our model, we used both geometric features, *i.e.*, the registered facial points (feature set I), and appearance features, *i.e.*, local binary patterns (LBP) histograms [61] (feature set II) extracted around each facial point from a region of 32×32 pixels. We chose these features as they showed good performance in variety of AU recognition tasks [10]. To reduce the dimensionality of the extracted features we applied PCA, retaining 95% of the energy. This resulted in approximately 20D (geometric) and 40D (appearance) feature vectors, for each dataset.

2) Evaluation Procedure: Some AUs occur rarely (*e.g.*, AU9,11,26 in CK+). Others do not exhibit strong co-occurrence patterns (*e.g.*, AU5 in DISFA). Hence, we selected the following subsets of highly correlated AUs: AUs (1, 2, 4, 6, 7, 12, 15, 17) for CK+, AUs (1, 2, 4, 6, 12, 15, 17) for DISFA and AUs (4, 6, 7, 9, 10, 43) for Shoulder-pain. The selected AUs occur jointly in the context of recorded expressions (*e.g.*, pain expression, see [58]). In order to prove the model's ability to deal with large number of outputs, we also show the performance when all AUs (from CK+) are used. A detailed description of the AUs used for the model evaluation is shown in Table I. We report the F1 score and the area under the ROC curve (AUC) as the performance measures. Both metrics are widely used in the literature as they quantify different characteristics of the classifier's performance. Specifically, F1, defined as $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$, is the harmonic mean between the precision and recall. It puts emphasis on the classification task, while being largely robust to imbalanced data (such as examples of different AUs). AUC quantifies the relation between true and false positives, showing the robustness of a classifier to the choice of its decision threshold. In all our experiments, we performed a 5 fold subject independent cross-validation.

3) Models Compared: We compare the proposed MC-LVM to GP methods with different learning strategy. Specifically, we compare to the manifold relevance determination (MRD) [47], which uses the variational approximation, to the discriminative shared Gaussian process latent variable model (DS-GPVLM) [57] and to the multi-task latent GP (MT-LGP) [46], which perform exact ML learning. We also compare to the multi-label backpropagation and kNN (k=1), *i.e.*, the BPMLL [39] and ML-KNN [38], respectively. Lastly, we compare to the state-of-the-art

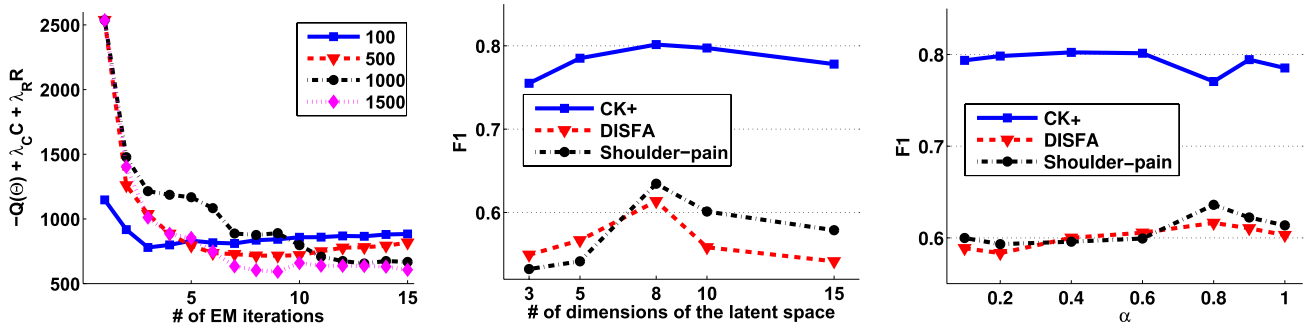


Fig. 6. The penalized negative log-likelihood of the MC-LVM for different number of samples used to estimate the posterior of the latent space (left), and average F1 score for multiple AU detection as a function of the dimensionality of the latent space (middle), and the regularization parameter α (right).

methods for multiple AU detection: the parametric methods Bayesian group-sparse compressed sensing (BGCS) [13], hierarchical RBM (HRBM) [15], joint patch multi-label learning (JPML) [14], and the kernel method l_p-regularized multi-task MKL (l_p-MTMKL) [17]. All the compared methods are evaluated using the same previously described input features. Note that implementation of JPML [14] was not available, and thus, in our comparison we report the results from the corresponding paper ([14] employed the SIFT appearance descriptor). For the single input methods (*i.e.*, BGCS, HRBM, BPMLL and ML-KNN), we concatenated the two feature sets. For the kernel-based methods, we used the RBF kernel. For l_p-MTMKL we also used the polynomial kernel, as suggested in [17]. Due to the high learning complexity of l_p-MTMKL ($\mathcal{O}(N^2T^2)$), where T is the number of target AUs), we followed the training scheme in [17] where multiple AUs were split into groups: $\{\{AU1, AU2, AU4\}, \{AU6, AU7, AU12\}, \{AU15, AU17\}\}$ for CK+, the same groups (without AU7) for DISFA, and $\{\{AU4, AU43, AU7\}, \{AU6, AU9, AU10\}\}$ for Shoulder-pain. The parameters of each method were tuned as described in the corresponding papers. For the MC-LVM, optimal values for the weighting parameters α, the regularization parameters λ_C, λ_R, as well as the size of the latent space were found via a validation procedure on the training set.

C. MC-LVM: Theoretical Evaluation

This section analyzes MC-LVM performance in terms of different parameter choices and settings. Fig. 6 (left) shows the convergence of the learning criterion in MC-LVM as a function of the used Monte Carlo samples during training on the CK+ dataset. We see that for small number of samples, the model does not converge to a (local) minimum. This is expected, since with $100-500$ samples the posterior in Eq. (3) cannot be approximated well. The model converges when 1000 samples are used, and its convergence does not change considerably after that. Thus, we fixed the number of samples to 1000. From Fig. 6 (middle), we see how the size of the latent space affects the performance of the learned model. It is clear that for both posed and spontaneous data, an 8-dimensional latent space is sufficient for the task of joint feature fusion and multiple AU detection, and results in the best average F1-score. Lower dimensional manifolds fail to explain the correlations between the input features and to capture the dependencies among multiple AUs, while manifolds with more than 8D do not include any additional discriminative information. Hence, in what follows, we fixed the size of the latent space to 8D. Fig. 6 (right) shows the effect of changing α on the discriminative power of the model. We observe that the model prefers a weighted conditional distribution over a fully generative or discriminative component. The optimal value of α is around 0.4 for posed, and 0.8 for spontaneous data. This difference is due to the fact that in case of spontaneous data (DISFA, Shoulder-pain), the model puts less focus on explaining unnecessary variations for the AU detection task, *e.g.*, due to the subject-specific features and errors due to the pose registration. Therefore, the influence of the generative component is lower (higher α) than in the case of posed facial expressions from CK+. Moreover, the CK+ dataset contains significantly less data (around 600 annotated frames) than DISFA and Shoulder-pain. Hence, MC-LVM prioritizes the generative component, to avoid overfitting the training data. On the other hand, when we have sufficient training examples (DISFA, Shoulder-pain), MC-LVM prefers to give less emphasis to the conditional distribution of the features (generative component). Such behavior of multi-conditional models has been also observed in other domains (e.g., in [22] for pixel classification).

To provide a better insight regarding the advantages of selecting a weighted conditional distribution, in Fig. 7 we compare the performance of the MC-LVM when the likelihood term consists of only the discriminative conditional ($\alpha = 1$), and the optimal weighted conditional ($\alpha = 0.4$ for CK+ and $\alpha = 0.8$ for DISFA and Shoulder-pain). We can see that the weighted conditional improves the performance on most of AUs, with significant enhancement in the performance on certain AUs (3% on AU7,15 on CK+, 6% on AU1 and 3% on AU6,15 on DISFA, and 10% on AU7,9,10 on Shoulder-pain).

In Fig. 8 (left) we see the effect of the introduced relational constraints on the model's performance. At first we observe that when no regularization is used ($\lambda_C, \lambda_R = 0$), MC-LVM achieves the lowest performance for both posed and spontaneous data. By including only the topological constraint ($\lambda_C \neq 0, \lambda_R = 0$), MC-LVM attains a better representation of the data in the manifold, which results in higher F1 scores. Finally, with the addition of the global relational constraint

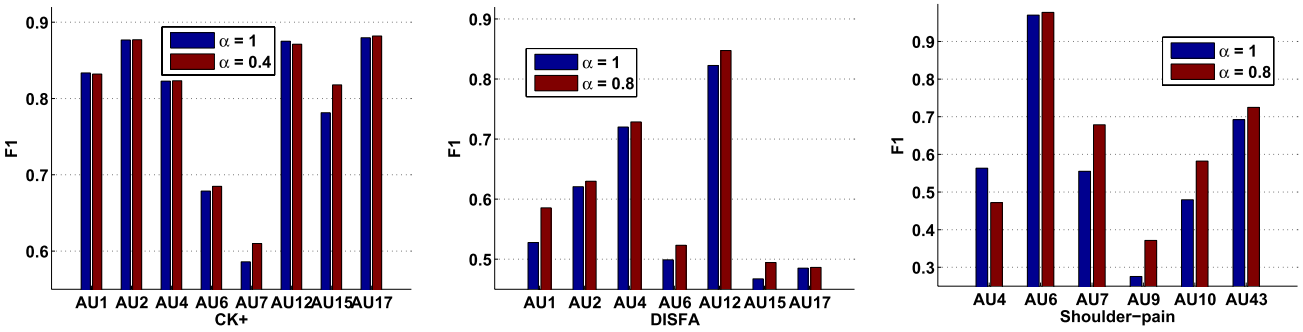


Fig. 7. Joint AU detection with MC-LVM on CK+ (left), DISFA (middle) and Shoulder-pain (right) for different value of α. The comparisons are between the discriminative-only conditional ($\alpha = 1$) and the optimal weighted conditionals ($\alpha = 0.4$ for CK+ and $\alpha = 0.8$ for DISFA and Shoulder-pain).

TABLE II

F1 SCORE AND AUC FOR JOINT AU DETECTION ON CK+ DATASET. COMPARISONS TO STATE-OF-THE-ART

Methods (I+II)	F1 score									AUC								
	AU1	AU2	AU4	AU6	AU7	AU12	AU15	AU17	Avg.	AU1	AU2	AU4	AU6	AU7	AU12	AU15	AU17	Avg.
MC-LVM	84.39	86.55	81.60	68.42	61.67	**88.48**	**82.54**	**87.40**	**80.14**	95.66	96.80	**93.97**	92.07	87.84	**97.78**	**94.60**	**96.10**	**94.35**
MC-LVM (SO)	86.06	**88.37**	**82.93**	70.80	57.27	87.16	73.26	85.57	78.93	**98.22**	97.25	93.95	92.20	85.71	97.41	94.05	95.80	94.33
MRD [47]	80.72	79.18	69.93	69.81	53.24	77.83	65.70	85.20	72.70	95.58	92.53	91.85	**92.73**	82.69	94.50	91.32	94.78	92.00
MT-LGP [46]	**89.12**	83.70	79.79	67.16	60.89	80.53	64.63	85.97	76.47	96.70	**97.33**	90.90	91.45	86.37	96.92	94.25	94.80	93.59
DS-GPLVM [57]	87.41	81.78	79.70	68.48	63.29	81.04	60.33	84.29	76.17	96.10	96.69	89.56	89.83	85.91	95.69	92.56	94.03	92.55
BGCS [13]	84.57	86.19	81.17	69.82	59.48	87.77	74.77	84.84	78.58	97.76	96.63	93.21	91.59	85.06	97.69	94.04	95.43	93.85
HRBM [15]	87.62	84.00	74.10	62.90	50.74	82.38	66.06	84.56	74.04	95.99	95.13	88.00	88.37	78.09	93.73	93.49	95.60	91.05
l_p-MTMKL [17]	87.50	85.50	51.43	**72.65**	58.82	85.95	74.21	75.44	73.93	93.19	94.99	90.95	90.01	84.41	95.67	91.06	92.97	91.65
BPMLL [39]	75.41	84.31	64.85	69.14	**64.34**	83.98	69.50	76.25	73.47	89.06	95.21	76.88	90.53	85.51	95.48	90.20	88.19	88.88
ML-KNN [38]	76.83	84.34	63.28	67.23	53.19	82.88	65.88	78.71	71.54	89.07	95.54	76.46	90.58	**90.71**	94.31	92.65	89.13	89.81
JPML* [14]	91.2	96.5	-	75.6	50.9	80.4	76.8	80.1	78.8	-	-	-	-	-	-	-	-	-

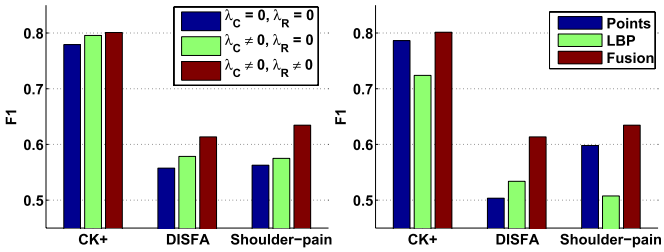


Fig. 8. Average F1 score on all three datasets. The effect of the relational constraints (left), and the feature fusion (right) on the joint AU detection task.

($\lambda_C, \lambda_R \neq 0$) MC-LVM achieves the highest scores. Note that the difference is more pronounced in data from DISFA and Shoulder-pain. This evidences the importance of modeling the global relations for the detection of spontaneous (more subtle) AUs. This is because the features of these AUs are corrupted by higher noise levels and thus, their joint prediction can help to reduce uncertainty of the classifiers, as has been reported in [62]. Fig. 8 (right) shows the average performance of the model for different feature combinations. In the single input case, we observe that, on average, geometric features (I) outperform the appearance features (II) (apart from DISFA where features (I) suffer from residual errors from the pose registration due to large variations in the head pose). This is because, by concatenating the LBP histograms obtained from each patch, the local information of the data is lost, and thus, the model obtains lower scores. However, when both inputs are used, MC-LVM can unravel a very informative shared latent space. This results in the highest F1 score, with significant improvement on the spontaneous data of DISFA and Shoulder-pain. In general, from Fig. 8 we see that the effect of the introduced regularization and the feature fusion is far more pronounced in the spontaneous expressions, where a limited and imbalanced number of examples is available for each AU.

D. Model Comparisons on Posed Data

We next compare the proposed MC-LVM to several state-of-the-art methods on the posed data from CK+. We first inspect the performance of MC-LVM and the GP-related methods. From Table II, we can see that the MAP-based methods, *i.e.*, the MT-LGP [46] and DS-GPLVM [57], achieve similar performance on average since they are based on the same learning scheme. On the other hand, MRD [47], uses a variational distribution to approximate a manifold shared across multiple inputs and outputs, without any additional constraints over the latent variables. This results in a poor accuracy. Also, MRD learns an approximation to the posterior, in order to predict the variational latent positions that best generate the inputs, while MT-LGP and DS-GPLVM learn accurate back mappings from the input spaces to the manifold. By contrast, the combination of the approximate learning with the relational constraints used in the proposed MC-LVM results in a significant increase in performance over the GP-based methods. We partly attribute this to the explicit modeling

TABLE III
F1 SCORE FOR JOINT AU DETECTION (ALL 17) ON CK+ DATASET. COMPARISON TO STATE-OF-THE-ART

Methods (I+II)	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU11	AU12	AU15	AU17	AU20	AU23	AU24	AU25	AU26	AU27	Avg.
MC-LVM	82.49	**86.96**	**79.16**	**73.47**	**72.80**	**57.52**	87.94	**31.11**	**87.60**	**76.40**	**86.76**	70.27	67.27	**51.02**	91.81	21.05	**91.14**	**71.45**
BGCS [13]	83.04	85.10	77.45	72.21	69.26	55.94	**89.03**	29.41	86.79	74.92	83.33	**71.10**	**68.01**	48.14	76.60	34.21	88.55	70.12
HRBM [15]	**86.86**	85.47	72.58	72.04	61.74	54.47	85.91	26.51	72.65	72.53	81.66	47.46	56.64	35.29	**92.57**	**37.61**	87.65	66.45

TABLE IV
F1 SCORE AND AUC FOR JOINT AU DETECTION ON DISFA DATASET. COMPARISONS TO THE STATE-OF-THE-ART

Methods (I+II)	F1 score								AUC							
	AU1	AU2	AU4	AU6	AU12	AU15	AU17	Avg.	AU1	AU2	AU4	AU6	AU12	AU15	AU17	Avg.
MC-LVM	**58.55**	**62.99**	**72.85**	52.32	**84.74**	**49.44**	**48.63**	**61.36**	**79.58**	84.01	**84.87**	62.75	**92.43**	78.97	**73.87**	**79.50**
MC-LVM (SO)	35.50	52.68	70.99	54.67	82.58	37.11	47.76	54.47	64.71	85.21	82.52	**68.15**	92.20	**79.22**	72.39	77.77
MT-LGP [46]	41.44	36.84	61.19	45.98	49.78	40.12	43.01	45.48	69.28	79.31	74.23	62.08	70.22	58.61	67.69	68.27
BGCS [13]	50.13	36.49	72.05	59.64	78.47	39.93	40.29	53.86	69.54	49.72	78.93	66.76	86.55	73.67	63.36	69.79
HRBM [15]	39.67	55.92	61.56	54.01	79.16	38.72	38.82	52.55	61.55	**85.88**	67.10	58.08	81.74	64.93	64.41	69.10
l_p-MTMKL [17]	42.21	45.81	47.18	**62.79**	76.33	34.47	41.40	50.03	71.77	73.42	62.49	66.27	78.83	59.16	63.98	67.98

of AU co-occurrences through the introduced constraints, as well as to the multi-conditional learning based on the proposed sampling scheme. The importance of the latter is further evidenced in the performance of the single output instance of MC-LVM, which for the case of the posed data achieves comparable scores to the multi-output. We see that joint learning does not improve detection of all AUs. It even shows reduced performance for certain AUs. For example, from Fig. 5, we see that AU1,2 are strongly correlated, yet single output achieves higher F1 on both AUs compared to the multi-output setting. This shows that for given data, these two AUs can be predicted well without relying on each other. On the other hand, the performance of AU15, which is strongly correlated with AU17, and has significantly less examples than other AUs, is considerably improved (F1 9% higher). The similar performance between the two settings is also explained from the nature of the posed data of CK+. Joint AU learning is expected to be advantageous, in cases where the input data suffer from high-dimensional noise [62]. Hence the superior performance of the multi-output setting will be evidenced in the evaluations on the spontaneous data from DISFA and Shoulder-pain in Sec. IV-E.

Table II, also summarizes the performance of the state-of-the-art models for joint AU detection: BGCS, HRBM and l_p-MTMKL. These models, manage to improve the detection of AU1 and AU6, by successfully modeling their co-occurrences between the related AUs (AU2 and AU12 respectively) in the expressions of Surprise and Happiness. However, their performance on more subtle AUs, *e.g.*, AU7,15,17 is significantly lower than that of the proposed MC-LVM. This is due to the fact that the parametric models BGCS and HRBM cannot handle simultaneously the fusion of the *concatenated* features and the modeling of the AU dependencies using compressed/binary latent variables. On the other hand, l_p-MTMKL can perform the fusion through the MKL framework. However, due to its modeling complexity, it is trained on subsets of AUs, which affects its ability to capture all AU relations. More importantly, in contrast to MC-LVM, these models lack the generative component, which, evidently, acts as a powerful regularizer. The results of JPML were obtained from [14], thus, they are not directly comparable to the other models. Yet, we report this performance as a reference to the state-of-the-art. Finally, the baseline multi-label methods, BPMLL and ML-KNN attempt to model the AU dependencies directly in the classifier level, as in l_p-MTMKL, but they cannot perform the fusion of the input features. Hence, they achieve the lowest average scores.

To demonstrate the model's scalability when dealing with large number of outputs, we compare the proposed approach to the state-of-the-art HRBM and BGCS for joint AU detection on *all* 17 AUs from CK+ (l_p-MTMKL cannot be evaluated on this experiment due to its learning complexity). As we can see from Table III, modeling the remaining (less frequent) AUs affects the overall performance of all three models, *i.e.*, MC-LVM, BGCS and HRBM, which suffer a drop of 8.6%, 8.4% and 7.6%, respectively. However, MC-LVM outperforms HRBM on 14 out of 17 AUs and BGCS on 12 out of 17 AUs, which demonstrates the ability of the former to better model the relations among AUs, even in case of many AU classes.

E. Model Comparisons on Spontaneous Data

We further investigate the models' performance on spontaneous data from DISFA and Shoulder-pain datasets. We focus here on the best performing methods from Table II. From Tables IV–V, we can observe a significant drop in the performance of all methods on both datasets. This evidences the difficulty of the task of AU detection in realistic environments, where spontaneous expressions are present. Also, typical for naturalistic data, the distribution of the activated AUs is more imbalanced than in the case of the posed dataset. This poses an additional modeling challenge since training data for certain AUs (*e.g.*, AU2,15 for DISFA, and AU9,10 for Shoulder-pain) are limited. Consequently, the models need to put more emphasis on the AU co-occurrences for detection of these AUs. As evidenced by the results in Tables IV–V, this adversely affects the single output MC-LVM. Contrary to the high achieved performance on the posed data, the single output instance reports here significantly lower scores

TABLE V

F1 SCORE AND AUC FOR JOINT AU DETECTION ON SHOULDER-PAIN DATASET. COMPARISONS TO THE STATE-OF-THE-ART

Methods (I+II)	F1 score							AUC						
	AU4	AU6	AU7	AU9	AU10	AU43	Avg.	AU4	AU6	AU7	AU9	AU10	AU43	Avg.
MC-LVM	47.20	**97.75**	67.88	37.13	58.23	**72.51**	**63.45**	53.58	**82.27**	57.80	54.65	87.80	66.13	**67.04**
MC-LVM (SO)	57.76	95.57	63.59	34.54	49.93	64.49	60.98	**66.36**	50.47	60.04	53.23	64.20	65.81	60.02
MT-LGP [46]	50.42	50.48	63.52	33.38	**61.62**	61.00	53.40	61.35	44.40	60.96	52.47	**90.39**	60.90	61.75
BGCS [13]	**61.42**	71.52	60.40	**37.86**	54.50	63.49	58.20	63.28	59.29	59.93	**59.23**	69.96	67.10	63.13
HRBM [15]	47.20	93.93	63.67	29.80	52.39	69.54	59.42	57.33	77.41	62.56	53.21	71.36	**73.19**	65.85
l_p-MTMKL [17]	37.69	**97.75**	**70.08**	33.28	41.79	44.03	54.10	54.95	71.86	**64.15**	53.84	68.62	64.69	63.01

TABLE VI

CROSS-DATASET EVALUATIONS OF THE STATE-OF-THE-ART MODELS ON 7 AUS PRESENT IN BOTH CK+ AND DISFA DATASETS. THE MODELS ARE TRAINED ON DATA FROM CK+ DATASET AND TESTED ON DATA FROM DISFA DATASET (C→D), AND THE OTHER WAY AROUND (D→C)

Train→Test	Methods (I+II)	F1 score								AUC							
		AU1	AU2	AU4	AU6	AU12	AU15	AU17	Avg.	AU1	AU2	AU4	AU6	AU12	AU15	AU17	Avg.
C→D	MC-LVM	53.92	**54.69**	**68.37**	51.99	70.77	**37.14**	42.81	54.24	76.78	86.80	**79.74**	**73.21**	86.73	62.28	67.83	76.20
	BGCS [13]	**59.01**	49.37	68.34	**57.75**	**80.26**	36.59	43.54	**56.41**	**86.75**	91.75	78.97	69.97	**87.83**	64.83	69.67	**78.54**
	HRBM [15]	43.20	36.83	52.10	36.15	40.70	35.61	**51.13**	42.25	67.41	71.84	65.62	59.32	62.62	60.77	**74.05**	65.95
	l_p-MTMKL [17]	39.13	41.24	44.77	49.42	69.67	31.55	39.12	44.98	71.77	73.42	72.70	68.38	67.46	69.31	65.85	65.56
D→C	MC-LVM	**72.22**	**85.85**	**75.05**	59.94	63.45	54.81	**73.35**	**69.24**	**92.51**	**96.60**	**90.51**	84.24	**95.02**	87.21	90.82	**90.99**
	BGCS [13]	61.11	71.90	67.84	**65.05**	**80.46**	54.23	69.98	67.22	84.44	91.21	88.21	84.91	94.54	84.12	84.97	87.49
	HRBM [15]	66.81	64.52	60.12	54.11	65.60	**60.47**	66.67	62.61	88.88	92.26	81.47	**88.23**	94.19	**87.91**	**91.61**	89.22
	l_p-MTMKL [17]	68.10	61.94	56.06	57.86	66.26	43.30	63.66	59.60	80.21	82.41	69.45	79.59	86.28	74.64	78.88	78.78

for the aforementioned AUs in both datasets. Furthermore, the small amount of training data for some AUs, imposes an additional difficulty when modeling the global AU relations. Consequently, the parametric discriminative models, BGCS and HRBM, overfit the data and report low performance. This exemplifies the importance of modeling the relations among the features via the generative component, in the proposed approach. Note that for some AUs with sufficient training data, *e.g.*, AU4,6 in DISFA, BGCS and HRBM achieve similar or better scores than the MC-LVM. This is in part due to modeling the multiple AU detectors under a joint cost function – each method selects to put more emphasis on modeling different AUs than the others. However, the MC-LVM outperforms these models on average. l_p-MTMKL obtains very low scores (especially in the Shoulder-pain), which is a result of not modeling global relations, due to its training scheme. MT-LGP also fails to model explicitly the relations between AUs, achieving low scores as well. The proposed MC-LVM is more robust to the data imbalance, and can better discover the AU relations, which in turn gives not only the best average F1 scores, but also achieves more robust performance as evidenced by the higher AUC.

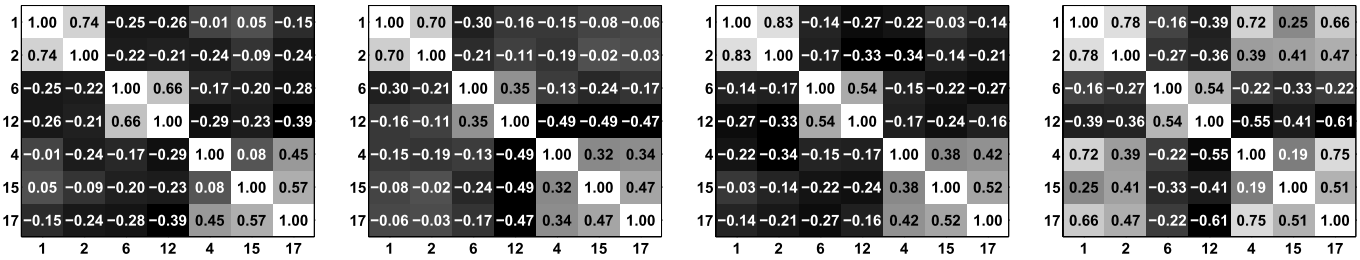
F. Cross Dataset Experiments on CK+ and DISFA

Herein, we evaluate the robustness of the models in a cross dataset experiment. Specifically, we perform two different cross-dataset experiments, CK+→DISFA and DISFA→CK+.[2] We evaluate the models' performance on 7 AUs (*i.e.*, $1, 2, 4, 6, 12, 15, 17$) that are present in both datasets. This is a rather challenging task due to the different characteristics of the data. First of all, as shown in Fig. 4, the facial images differ in terms of illumination, pose and size, which imposes a further difficulty on the alignment of the input facial features. Another key challenge is the difference in the context of the two datasets. The data from CK+ contain posed expressions, which vary considerably in subtlety compared to the spontaneous data of DISFA. The latter also affects the co-occurrence patterns among the AUs, as can be seen from Fig. 5

From Table VI, we see that the performance of the models is lower for most of AUs compared to that attained on the original dataset (see Tables II-V). This is expected for the reasons mentioned above. Interestingly, BGCS achieves higher performance on the cross dataset experiment CK+→DISFA, than when both training and testing is performed on DISFA dataset. This confirms our claims in Section IV-E that this method cannot fully unravel the dependencies among the AUs when dealing with imbalanced data in the training phase. The parametric model, *i.e.*, BGCS, can better model the AU relations with small (but well distributed) amount of training data, as in CK+. Hence, it achieves higher performance compared to MC-LVM. However, on the DISFA→CK+ experiment, we see that the proposed MC-LVM, benefits from the use of the non-parametric feature fusion, and manages to successfully unravel the structure and the co-occurrence patterns in the data, regardless of the imbalances in the amount of training examples and the subtlety of the spontaneous facial expressions. Thus, it attains superior performance compared to the BGCS, especially for AU1,2,4, where the two models achieve similar predictions for training and testing on CK+ (see Table II). Finally, the proposed MC-LVM consistently outperforms HRBM and l_p-MTMKL on both cross-dataset experiments, as evidenced from both F1 and AUC results.

Finally, in Fig. 9 we see the recovered AU dependencies from the MC-LVM, on the test data in both within and cross-dataset experiments. As we observe from Fig. 9(a)&(c),

[2] 'A→B' denotes the training on dataset A and testing on dataset B.



(a) CK+→CK+ (b) DISFA→DISFA (c) DISFA→CK+ (d) CK+→DISFA

Fig. 9. The learned global AU relations (in terms of correlation coefficients) for within datasets (a),(b) and cross-datasets (c),(d) experiments.

the recovered AU dependencies for CK+ are similar to the original co-occurrence patterns from Fig. 5. Hence, the proposed MC-LVM attains competitive results for CK+ and the DISFA→CK+ experiments. On the other hand, by comparing Fig. 9(d)&(b) and Fig. 5, we observe that MC-LVM has falsely recovered strong correlations between AU1,2 and AU15,17, which results in the low performance in the CK+→DISFA experiment. We attribute this to the fact that AU1,4,17 are the dominant AUs in CK+, which is not the case for DISFA. Thus the model trained on CK+ seems to have a bias on predicting AU1,4,17. Due to their strong relations with AU2,15 MC-LVM recovers the false dependencies on DISFA dataset.

V. Conclusions

To conclude, we proposed the multi-conditional latent variable model that brings together GPs and multi-conditional learning to achieve a feature fusion for multi-label classification of facial AUs. The majority of existing approaches perform feature fusion via simple vector concatenation. However, this leads to the false assumption that the multiple feature sets are identically distributed. By assuming conditional independence given the subspace of AUs, MC-LVM learns different distributions for each feature set via separate GPs, resulting in more accurate fusion in the manifold, and hence, more discriminative features for the detection task. More importantly, the newly introduced multi-conditional objective allows the generative and discriminative costs to act in concert during the model learning – the generative component has the key role in unraveling the latent space for the feature fusion, while the discriminative component endows the space with the relational/class information of the outputs. The retrieved manifold leads to superior performance compared to other solely discriminative or generative approaches. We further proved that the novel *topological* and *relational* constraints can increase the discriminative power of the model, by successfully encoding the AU dependencies into the learned manifold. We demonstrated the effectiveness of these properties on three publicly available datasets, and showed that the proposed model outperforms the existing works for multiple AU detection, and several methods for feature fusion and multi-label learning. We also showed that the proposed model is able to generalize across different datasets.

One main limitation of the proposed approach is its inefficiency to deal with large data during training. As purely based on the framework of GPs, MC-LVM's training scales in $\mathcal{O}(N^3)$, which typically imposes a restriction on using datasets of size $\mathcal{O}(10^4)$. However, this can be addressed by sparse [63] or distributed [64] computations, which scale GPs to $\mathcal{O}(10^7)$. An extra burden during the training of MC-LVM is the requirement for manual selection of the weighting between the generative and discriminative components. Ideally, within our probabilistic formulation, the balancing of the conditional distributions should be handled automatically. Finally, as evidenced by our experiments, the proposed joint inference improves detection of most AUs and the overall performance. Yet, sometimes this results in decreased detection performance on other AUs, when compared to single output AU detectors. It would be interesting to investigate how the subsets of strongly correlated AUs could efficiently be detangled by learning subset-specific subspaces within the proposed framework. All these are possible directions of future work.

References

[1] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *APA Psychol. Bull.*, vol. 111, no. 2, p. 256–274, 1992.

[2] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Phil. Trans. Roy. Soc. B*, vol. 364, no. 1535, pp. 3505–3513, Dec. 2009.

[3] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*. Salt Lake City, UT, USA: Consulting Psychologists Press, 2002.

[4] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[5] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 28–43, Feb. 2012.

[6] K. R. Scherer and P. Ekman, *New Handbook of Methods in Nonverbal Behavior Research* (Series in Affective Science), 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 1982.

[7] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 568–573.

[8] W.-S. Chu, F. D. L. Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3515–3522.

[9] P. Lucey *et al.*, "Automatically detecting pain in video through facial action units," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 41, no. 3, pp. 664–674, Jun. 2011.

[10] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial action recognition combining heterogeneous features via multikernel learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 993–1005, Aug. 2012.

[11] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 336–342.

[12] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.

[13] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor, "Exploiting sparsity and co-occurrence structure for action unit recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2015, pp. 1–8.

[14] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2207–2216.

[15] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3304–3311.

[16] Y. Zhu, S. Wang, L. Yue, and Q. Ji, "Multiple-facial action unit recognition by shared feature learning and semantic relation modeling," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 1663–1668.

[17] X. Zhang, M. H. Mahoor, S. M. Mavadati, and J. F. Cohn, "A l_p-norm MTMKL framework for simultaneous detection of multiple facial action units," in *Proc. IEEE Int. Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 1104–1111.

[18] A. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao, "Learning shared latent structure for image synthesis and robotic imitation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 18. 2006, pp. 1233–1240.

[19] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[20] J. Zhang, Z. Ghahramani, and Y. Yang, "Flexible latent variable models for multi-task learning," *Mach. Learn.*, vol. 73, no. 3, pp. 221–242, 2008.

[21] A. Agarwal, S. Gerber, and H. Daume, "Learning multiple tasks using manifold regularization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 46–54.

[22] B. M. Kelm, C. Pal, and A. McCallum, "Combining generative and discriminative methods for pixel classification with multi-conditional learning," in *Proc. Int. Conf. Pattern Recognit.*, 2006, pp. 828–832.

[23] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Multi-conditional latent variable model for joint facial action unit detection," in *Proc. IEEE Int.Conf. Comput. Vis.*, 2015, pp. 3792–3800.

[24] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1940–1954, Nov. 2010.

[25] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 944–958, May 2015.

[26] O. Rudovic, V. Pavlovic, and M. Pantic, "Kernel conditional ordinal random fields for temporal segmentation of facial action units," in *Proc. Eur. Conf. Comput. Vis., Workshops*, 2012, pp. 260–269.

[27] S. Kaltwang, S. Todorovic, and M. Pantic, "Latent trees for estimating intensity of facial action units," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 296–304.

[28] 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 Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[29] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *Proc. Int. Symp. Vis. Comput.*, 2012, pp. 368–377.

[30] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, Jul. 2011.

[31] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 24–38, Jan. 2016.

[32] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 113–126, Jan. 2014.

[33] S. Z. Li, Z. Lei, and M. Ao, "The HFB face database for heterogeneous face biometrics research," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Workshops*, Jun. 2009, pp. 1–8.

[34] A. K. Jain, B. Klare, and U. Park, "Face matching and retrieval in forensics applications," *IEEE Multimedia*, vol. 19, no. 1, pp. 1–9, Jan. 2012.

[35] N. M. Correa, Y.-O. Li, T. Adali, and V. D. Calhoun, "Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 385–388.

[36] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[37] M. S. Sorower, *A Literature Survey on Algorithms for Multi-Label Learning*. Corvallis, OR, USA: Oregon State Univ., 2010.

[38] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

[39] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.

[40] T. Finley and T. Joachims, "Training structural SVMs when exact inference is intractable," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 304–311.

[41] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proc. Int. Conf. Mach. Learn.*, 2004, p. 104.

[42] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 109–117.

[43] A. Kumar and H. Daume, III, "Learning task grouping and overlap in multi-task learning," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1383–1390.

[44] T. V. Nguyen and E. V. Bonilla, "Collaborative multi-output Gaussian processes," in *Proc. Int. Conf. Uncertainty Artif. Intell.*, 2014, pp. 643–652.

[45] A. Argyriou, S. Clémençon, and R. Zhang, "Learning the graph of relations among multiple tasks," GALEN-INRIA Saclay, France, Tech. Rep. HAL-00940321, 2013.

[46] R. Urtasun, A. Quattoni, N. D. Lawrence, and T. Darrell, "Transferring nonlinear representations using Gaussian processes with a shared latent space," MIT, Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-08-020, 2008.

[47] A. Damianou, C. H. Ek, M. Titsias, and N. Lawrence, "Manifold relevance determination," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 145–152.

[48] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.

[49] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[50] L. Bo and C. Sminchisescu, "Supervised spectral latent variable models," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2009, pp. 33–40.

[51] A. McCallum, C. Pal, G. Druck, and X. Wang, "Multi-conditional learning: Generative/discriminative training for clustering and classification," in *Proc. Amer. Assoc. Artif. Intell.*, vol. 21. 2006, p. 433–439.

[52] N. D. Lawrence and J. Q. Candela, "Local distance preservation in the GP-LVM through back constraints," in *Proc. Int. Conf. Mach. Learn.*, vol. 148. 2006, pp. 513–520.

[53] F. R. Chung, *Spectral Graph Theory*. Providence, RI, USA: AMS, 1997.

[54] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. London, U.K.: Chapman & Hall, 2005.

[55] X. He, D. Cai, Y. Shao, H. Bao, and J. Han, "Laplacian regularized Gaussian mixture model for data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 9, pp. 1406–1418, Sep. 2011.

[56] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *J. Mach. Learn. Res.*, vol. 6, pp. 1783–1816, Nov. 2005.

[57] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 189–204, Jan. 2015.

[58] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2011, pp. 57–64.

[59] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013.

[60] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, 2004.

[61] T. Ojala, M. Pietikáinen, and T. Máenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[62] A. Maurer, "Bounds for linear multi-task learning," *J. Mach. Learn. Res.*, vol. 7, pp. 117–139, Jan. 2006.

[63] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in *Proc. Adv. Neural Inf. Proccess. Syst.*, 2005, pp. 1257–1264.
[64] M. P. Deisenroth and J. W. Ng, "Distributed Gaussian processes," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–10.

Stefanos Eleftheriadis received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 2011. He is currently pursuing the Ph.D. degree with the Computing Department, Imperial College London, U.K. His research interests are in machine learning and computer vision with applications to automatic human behavior analysis. He received the National Award in Microsoft's Imagine Cup software development competition, in 2011.

Ognjen Rudovic (M'15) received the Ph.D. degree from the Computing Department, Imperial College London, U.K., in 2014, and the M.Sc. degree in computer vision and artificial intelligence from the Computer Vision Center, Spain, in 2008. He is currently a Research Associate with the Computing Department, Imperial College London, U.K. His research interests are in automatic recognition of human affect, machine learning, and computer vision.

Maja Pantic (M'98–SM'06–F'12) is currently a Professor in affective and behavioral computing with the Department of Computing, Imperial College London, U.K., and the Department of Computer Science, University of Twente, The Netherlands. She currently serves as the Editor-in-Chief of the *Image and Vision Computing Journal* and an Associate Editor for both the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING. She was a recipient of 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.