Multi-conditional Latent Variable Model for Joint Facial Action Unit Detection

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

Abstract

We propose a novel multi-conditional latent variable model for simultaneous facial feature fusion and detection of facial action units. In our approach we exploit the structure-discovery capabilities of generative models such as Gaussian processes, and the discriminative power of classifiers such as logistic function. This leads to superior performance compared to existing classifiers for the target task that exploit either the discriminative or generative property, but not both. The model learning is performed via an efficient, newly proposed Bayesian learning strategy based on Monte Carlo sampling. Consequently, the learned model is robust to data overfitting, regardless of the number of both input features and jointly estimated facial action units. Extensive qualitative and quantitative experimental evaluations are performed on three publicly available datasets (CK+, Shoulder-pain and DISFA). We show that the proposed model outperforms the state-of-the-art methods for the target task on (i) feature fusion, and (ii) multiple facial action unit detection.

1. 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 [14]. The facial action coding system (FACS) [7] 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 33 unique AUs, several categories of head/eye positions and other movements, which can describe every possible facial expression.

Automatic detection of AUs is a challenging task mainly due to the complexity and subtlety of human facial behavior, and individual differences and artifacts caused by variation in head-pose, illumination, occlusions, etc. [4]. These and other sources of variation in facial expression data are typically accounted for at the (i) feature level, by finding facial features that are robust to the aforementioned artifacts, and/or (ii) model level, by capturing semantics of AUs, *i.e.*, their co-occurrences as commonly encountered in naturalistic data. At the feature level, detection of AUs can be performed using either geometric or appearance descriptors, or both. While the geometric features (*e.g.*, the displacement of the facial points between expressive and neutral faces [13]) are more robust to illumination and pose changes, not all AUs can be detected solely from them. For example, activation of AU6 wrinkles the skin around the outer corners of the eyes and raises the cheeks, which makes it difficult to detect this AU (independently from other AUs) using the geometric features explicitly. On the other hand, appearance-based features overcome this by being able to capture transient differences in the facial texture, such as wrinkles, bulges and furrows, however, they are usually prone to overfitting. Hence, modeling both geometric and appearance features exploits the complementary properties of these two features, leading to improved AU detection.

At the model level, the goal is to improve the AU detection by modeling 'semantics' of facial behavior (e.g., in terms of AU co-occurrences). This is important because AUs rarely appear in isolation (more than 7,000 AU combinations have been observed in everyday life [23]). The type of the AU co-occurrences depends largely on the context in which the facial expression is displayed, *e.g.*, due to latent variables such as emotions (*e.g.*, AU12 and AU6 in the case of happiness, and AU4 and AU7 in the case of fear). Furthermore, co-occurring AUs can be non-additive, in the case of one AU masks another, or a new and distinct set of appearances can be created [7]. For instance, AU4 (brow lowerer) has a different appearance when occurring together with AU1 (inner brow raise) than alone. When AU1,4 co-occur, the brows are drawn together and are raised due to the action of AU1; the brows are lowered otherwise. This, in turn, significantly affects the appearance of the target AUs.

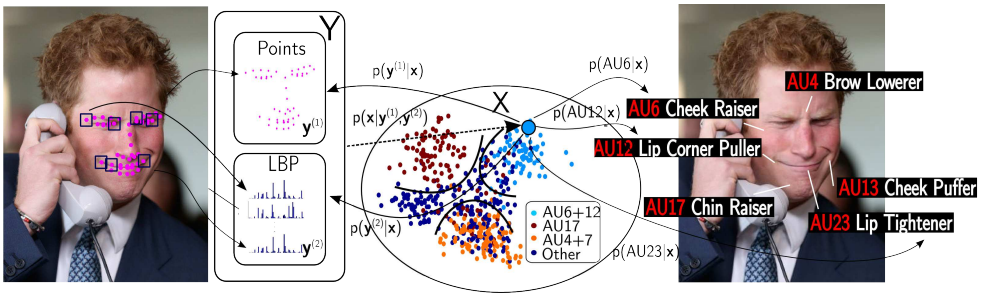


Figure 1. The proposed MC-LVM. The geometrical and appearance input features, $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$, are first projected onto the shared manifold $\mathbf{X}$. The fusion is attained via GP conditionals, $p(\mathbf{y}^{(1)}|\mathbf{x})$ and $p(\mathbf{y}^{(2)}|\mathbf{x})$, that generate the inputs. Classification is performed on the manifold via simultaneously learned logistic functions $p(z^{(c)}|\mathbf{x})$ 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.

Most of existing approaches to AU detection model each AU independently, using either a single feature set [2, 4], or by combining multiple feature sets through feature concatenation [13, 14], or multiple-kernel learning (MKL) [24]. Other methods treat different combinations of AUs as new independent classes [16]; yet, this is impractical given the number of possible combinations. On the other hand, methods that do attempt to model the AU co-occurrences (*e.g.*, [26, 33, 35]) fail to perform efficient fusion of different types of facial features. To the best of our knowledge, the only methods that attempt both are [29, 36, 34]. However, neither of these methods can perform simultaneous feature fusion and modeling of a large number of AUs.

To this end, we propose a Multi-Conditional Latent Variable Model (MC-LVM) that performs jointly the fusion of different facial features and detection of multiple AUs. Instead of performing the AU detection in the original feature space, as done in existing works [29, 34, 36], the MC-LVM attains the fusion by learning a low-dimensional subspace (i) shared across different feature sets, learned via the framework of Gaussian processes (GPs) [21], and (ii) constrained by the *local* dependencies among multiple AUs, encoded by means of string kernels [21], and the *global* dependencies, encoded via the AU co-occurrence structure. The *key* to our approach is the proposed definition of the multi-conditional likelihood function that combines both the generative and discriminative properties of probabilistic models. In contrast to existing subspace learning methods for multi-output (*e.g.*, [30]), the MC-LVM learns a discriminative subspace for multiple AU detection that is endowed with the generative property of GPs, which turns out to be an efficient regularizer during the parameter learning. To further improve the robustness of the parameter estimation, a Bayesian learning of the subspace is facilitated through Monte Carlo (MC) sampling, and the Expectation-Maximization (EM)-like algorithm is proposed. As a result, the training of the MC-LVM can be performed with a large number of AUs, without seriously affecting its computational load. During inference, multiple AU detection is performed through the learned subspace that best generate the input features. This is attained via the learned back mappings to the shared space, and does not require any additional optimization. As evidenced by our results, the resulting model achieves superior performance compared to existing methods for multiple AU detection, and other methods for feature fusion and multi-label classification. The outline of the proposed approach is given in Fig. 1.

2. Related Work

2.1. Facial AU detection

The majority of the existing works attempt to recognize AUs or certain AU combinations independently [13, 14, 2, 4, 16, 11, 22]. While the former ignores the dependencies among AUs, the latter is a prohibitively large space of possible combinations. To the best of our knowledge, there are only few works that perform joint AU detection. [26] proposed a generative framework based on dynamic Bayesian networks (DBN) to model the semantics of different AUs. A downside of this model, is that it lacks discriminative properties of what models. In contrast, the models in [36, 35, 29, 33, 34] are defined in a fully discriminative manner. Specifically, [36] 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. This independent modeling could result in inconsistent dependencies across inputs/outputs, and produce contradictory predictions. [35] 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. [29] employed the restricted Boltzman machine (RBM) to overcome the pair-wise AU modeling limitation of DBN. *Discrete* latent variables account for the dependencies among the outputs, which 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 is used for the fusion task. [33, 34] combine 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 features with different kernels [34], and account for the AU-dependencies [33]. Yet, [34] can deal only with subsets of AUs in its output due to its learning complexity, while in [33] the relations among the AUs are captured by *predefined* latent variables.

Our approach significantly differs from the above works, since the fusion of the features is performed in a continuous latent space. The latter can also efficiently model relations among large number of outputs, without the requirement to *a priori* define groups of AUs as done in [33, 34, 35]. The learning of the output dependencies is performed simultaneously to the fusion task by combining both generative and discriminative learning within a single model. This has not been addressed before in models for multiple AU detection.

2.2. Multi-label Classification

Various approaches for multi-label classification (MLC) exist in the literature. For an extensive overview please see [27, 25]. Baseline methods include [32], which extends the k-nearest neighbor (kNN) classifier to the multi-label scenario, and [31], which derives the back-propagation algorithm of the neural networks for the MLC. MLC is also highly related to multi-task learning techniques, that capture dependencies among multiple outputs through parameter sharing [9]. More sophisticated algorithms learn a latent variable model of task specific parameters within a probabilistic framework [30]. However, none of these methods perform simultaneous feature fusion and MLC.

To mitigate the limitations of the above methods, recent works in the GP context [28, 6] try to combine multi-task learning and feature fusion via subspace learning. [28] 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 the latent space is directly optimized using the ML strategy, which in the case of large number of data can overfit. To ameliorate this, [6] proposed learning of the space in a fully Bayesian framework using variational inference to integrate out the latent space.

Contrary to [28, 6], MC-LVM employs multi-conditional learning strategies to re-weight the generative and discriminative conditionals, in order to unravel a more suitable subspace for joint feature fusion and MLC. In our Bayesian approach, the latent space is approximated via efficient MC sampling, where the conditional models determine the importance of each sample. More importantly, the inference step is efficiently facilitated via the learned projection mappings to the manifold. This overcomes the requirement of [6] to learn another approximation to the posterior of the test inputs. Finally, note that such an approach has not been applied before on the task of multiple AU detection.

3. Multi-conditional Latent Variable Model (MC-LVM)

Let us denote the training set as $\mathcal{D} = \{\mathbf{Y}, \mathbf{Z}\}$. $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N]^T$ is comprised of N instances of multi-variate inputs stored in $\mathbf{y}_i = \{\mathbf{y}_i^{(1)}, \ldots \mathbf{y}_i^{(v)}, \ldots \mathbf{y}_i^{(V)}\}$, where $\mathbf{y}_i^{(v)} \in \mathcal{R}^{D_v}$. These represent different types of corresponding facial features or observation spaces. Furthermore, $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_N]^T$ are multiple binary labels, with $\mathbf{z}_i \in \{-1, +1\}^C$ encoding C (co-occurring) outputs.

3.1. Model Definition

We aim to learn a model that simultaneously combines different inputs and detects activations of multiple outputs. We assume the existence of a latent space $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N]^T$, where $\mathbf{x}_i \in \mathcal{R}^q$, $q \ll D$, jointly generates $\mathbf{y}_i$ and $\mathbf{z}_i$. For notational simplicity, in what follows, we set the number of input spaces to $V = 2$. Then, the joint distribution $p(\mathbf{y}, \mathbf{z})$ can formally be written down as marginalization over the latent space x as:

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

where we exploited the property of conditional independence, *i.e.*, $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{z}\}$ are independent given $\mathbf{x}$. For the non-linear conditional models, which we propose in Sec.3.2, the integral in Eq.(1) cannot be computed analytically. To this end, we numerically approximate the marginal likelihood using MC sampling

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

where the samples $\mathbf{x}_s$, $s = 1, \ldots, S$ are drawn from $p(\mathbf{x})$, which is defined in Sec.3.2. Using the Bayes' rule, we can derive the posterior of the model as:

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

We can now calculate the above probability for all pairs of training data i and MC latent samples s, to obtain the membership probabilities $p(s, i) = p(\mathbf{x}_s|\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \mathbf{z}_i)$. This gives rise to the expectation of the latent points:

$$\mathbf{x}_i = E\{\mathbf{x}|\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \mathbf{z}_i\} = \sum_{s=1}^{S} p(s, i) \mathbf{x}_s. \quad (4)$$

3.2. Conditional Models

The choice of conditional models $p(\mathbf{y}^{(v)}|\mathbf{x}), v = 1, 2$, and $p(\mathbf{z}|\mathbf{x})$, as well as the sampling distribution $p(\mathbf{x})$, in

Eq.(3) critically affect the representational capacity of the space, and thus, the model's performance. Effectively, this boils down to learning conditional models that provide: (i) *generative* mappings from latent space to the inputs ($\mathbf{x} \rightarrow \mathbf{y}^{(v)}, v = 1, 2$), (ii) *projection* mappings from the inputs to latent space ($\mathbf{y}^{(v)} \rightarrow \mathbf{x}$), and (iii), *discriminative* mappings from latent space to multiple binary outputs ($\mathbf{x} \rightarrow \mathbf{z}$).

Generative mappings. Different probabilistic models such as Gaussian models [3] or naive Bayes models [19] can be employed to recover the generative mappings. However, these parametric models are limited in their ability to recover non-linear mappings from the latent space to high-dimensional input features, and the other way around. To this end, we exploit the framework of GP [21], which allows us to model arbitrary data structures via suitable choice of a kernel function. We briefly describe GPs below.

Given a collection of latent points $\mathbf{X}$ and corresponding outputs, *e.g.*, $\mathbf{Y}^{(v)}$, we seek to find mapping $f : \mathbf{X} \rightarrow \mathbf{Y}^{(v)}$. By placing a GP prior over f, we can integrate it out [21]. Then, the marginal distribution over the outputs is:

$$p(\mathbf{Y}^{(v)}|\mathbf{X}, \theta_{Y^{(v)}}) = \frac{1}{\sqrt{(2\pi)^{ND_v}|\mathbf{K}_{Y^{(v)}}|^{D_v}}} \exp(-\frac{1}{2}\text{tr}((\mathbf{K}_{Y^{(v)}})^{-1}\mathbf{Y}^{(v)}(\mathbf{Y}^{(v)})^T)), \tag{5}$$

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

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

where $\delta_{\mathbf{x},\mathbf{x}'}$ is the Kronecker delta function, and $\theta_{Y^{(v)}} = (\theta_1, \theta_2, \theta_3, \theta_4)$ are the kernel hyperparameters. The parameter learning is performed by gradient-based minimization of $-log(p(\mathbf{Y}^{(v)}|\mathbf{X}, \theta_{Y^{(v)}}))$ w.r.t. $\theta_{Y^{(v)}}$ [21]. Then, conditional probability for new inputs x_* has the Gaussian form

$$p(\mathbf{y}_*^{(v)}|\mathbf{x}_*, \mathbf{X}, \mathbf{Y}^{(v)}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}_*^{(v)}}, \sigma_{\mathbf{y}_*^{(v)}}) \tag{7}$$

$$\boldsymbol{\mu}_{\mathbf{y}_*^{(v)}} = \mathbf{k}_*^T(\mathbf{K}_{\mathbf{Y}^{(v)}} + \sigma^2\mathbf{I})^{-1}\mathbf{Y}^{(v)} \tag{8}$$

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

The kernel values $\mathbf{k}_*$ and k_{**} are computed by applying Eq.(6) to $(\mathbf{X}, \mathbf{x}_*)$ and $(\mathbf{x}_*, \mathbf{x}_*)$, respectively, and σ is the noise on the outputs. We use the conditional model in Eq.(7) to represent $p(\mathbf{y}^{(v)}|\mathbf{x}), v = 1, 2$, in Eq.(3).

Projection mappings and sampling. To model the sampling distribution $p(\mathbf{x})$, the simplest choice is to assume a Gaussian prior over the latent points $\mathbf{x}$. However, sampling from such an uninformative prior, would give rise to latent representations that do not exploit the true nature of the input data. To ameliorate this, we define the sampling distribution so that it constraints the samples $\mathbf{x}_s$ by conditioning them on the inputs, *i.e.*, $\tilde{p}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)})$. This is motivated by the notion of back-constraints in [12], where this type of conditional distribution is used to learn the mappings from input to latent space, and also ensures that distances between the outputs (in our case, $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}\}$) are preserved in the manifold. We learn the conditional model for $\tilde{p}(\mathbf{x})$ using GPs, as done for the generative mappings. The use of GP in the projection mappings allows us to easily combine multiple features within its kernel matrix as $\mathbf{K}_X = \mathbf{K}_X^{(1)} + \mathbf{K}_X^{(2)}$, corresponding to the sum of the kernel functions defined on $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$, respectively. The resulting conditional model $p(\mathbf{x}_*|\mathbf{y}_*^{(1)}, \mathbf{y}_*^{(2)})$, which is the Gaussian distribution as in Eq.(7), is used for sampling.

Discriminative mappings. Since we are interested in detection of activations of multiple AUs, we use the logistic function [21] to model $p(\mathbf{z}|\mathbf{x})$. By assuming conditional independence given $\mathbf{x}$, we can factorize this conditional as:

$$p(\mathbf{z}|\mathbf{x}, \mathbf{W}) = p(z^{(1)}|\mathbf{x}, \mathbf{w}_1) \dots p(z^{(C)}|\mathbf{x}, \mathbf{w}_C), \tag{10}$$

$$p(z^{(c)}|\mathbf{x}, \mathbf{w}_c) = \frac{1}{1 + e^{-\mathbf{x}^T\mathbf{w}_c z^{(c)}}},\ c = 1, \dots, C, \tag{11}$$

where $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_C] \in \mathcal{R}^{q \times C}$ contains the weight vectors of the individual functions. During inference, if $p(z_*^{(c)}|\mathbf{x}_*) > 0.5$, the c-th output is active, *i.e.*, $z_*^{(c)} = 1$. In the case of multi-class outputs (*e.g.*, when modeling AU intensities), the class conditional in Eq.(11) should be modeled with multiple logistic functions.

3.3. Output Relational Constraints

Due to the possible large number of outputs, the topology of the latent space need to be constrained in order to avoid the model focusing on unimportant variation in the data. We need, also, to encourage the model to produce similar predictions for likely co-occurring outputs (*e.g.*, AU6+12), and dissimilar for some rarely co-occurring (*e.g.*, AU12 and AU17). Below we describe the construction of appropriate constraints based on the output relations, and how these are incorporated into the MC-LVM as additional regularizers.

Topological constraints. Herein, we define constraints that encode co-occurrences of the output labels using the notion of the graph Laplacian matrix [5]. The latter is defined as $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{S}$ is a $N \times N$ similarity matrix, and $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{S}_{ij}$. We define S in a supervised fashion by measuring the similarity between the output label vectors using string kernels [21] as:

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

where $\mathcal{A}$ is the set of all possible 2^C combinations of sub-labels l for a given latent position, and $\mathbf{z}_{l,x}$ denotes the number of times l appears in labels $\mathbf{z}$ of $\mathbf{x}$. Note that by accounting for all sub-labels, we measure the similarity of the outputs based on all possible groups of AUs, and not only on pairs. Then, using the expectation of the latent positions from Eq. (4), we arrive at the Laplacian regularization term:

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

Eq. (13) incurs higher penalty if latent projections of co-occurring AUs are distant in the latent space.

Global relational constraint. In order for the MC-LVM to fully benefit from the above topological constraint, it is important to ensure that the model will produce similar predictions for frequently co-occurring AUs. For this, we introduce the *global relational regularizer* as:

$$R = \|\mathbf{P}_\mathbf{z}^T\mathbf{P}_\mathbf{z} - \mathbf{Z}_0^T\mathbf{Z}_0\|_F^2, \tag{14}$$

where $\mathbf{P}_\mathbf{z} = [p(\mathbf{z}_1|\mathbf{x}_1), \ldots, p(\mathbf{z}_N|\mathbf{x}_N)]^T$ are the predictions from Eq.(11) for each $\mathbf{x}_i$ from Eq.(4), and $\mathbf{Z}_0$[1] is the true label set. Thus, the regularizer in Eq.(14), incurs a high penalty if correlated outputs have dissimilar predictions.

3.4. 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)

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

where $\mathbf{\Theta} = \{\theta_{Y^{(v)}}, \mathbf{W}\}$. Note that, in contrast to the standard ML optimization, we use the parameter $\alpha \in [0,1]$ to find an optimal balance between the generative ($p_{g,i}$) and discriminative ($p_{d,i}$) components, as commonly used in multi-conditional models [19]. The generative component has the key role to unravel the latent space of the fused features, while the discriminative component regularizes the manifold by inducing to the space information regarding the outputs' relations. By finding optimal α, we restructure the joint likelihood by allowing the model to concentrate its modeling power on a conditional distribution of interest.

To optimize the objective in Eq.(15), 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

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

[1]The subscript 0 indicates the negative class.

Algorithm 1 MC-LVM: Learning and Inference

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

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

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

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

w.r.t. $\mathbf{\Theta}$ using the conjugate gradient method [21].

The full training of the model is split into two stages, where in each stage we compute $p(\mathbf{x}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)})$ and $p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{z}|\mathbf{x})$ alternatively. First, we initialize the latent coordinates $\mathbf{X}$, using a dimensionality reduction method, *e.g.*, PCA. Then, we learn the sampling distribution $p(\mathbf{x}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)})$ by training a GP on the projection mappings, as explained in Sec.3.2, and collect S samples from the GP posterior. During the second stage, we employ the EM algorithm described above to learn the parameters $\mathbf{\Theta}$. Note that the constraints C and R implicitly depend on the posterior, which is a function of the current $\mathbf{\Theta}$, hence, we need to compute their derivatives w.r.t to $\mathbf{\Theta}$. Eq.(17) can be optimized jointly [3] or separately [10] 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 $\mathbf{X}$, using Eq.(4). We iterate between stage 1 and 2 until convergence of the objective function in Eq.(17).

Inference: Inference in MC-LVM is straightforward. The test data $\mathbf{y}_*^{(1)}, \mathbf{y}_*^{(2)}$, 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. All this is summarized in Alg. 1.

4. Experiments

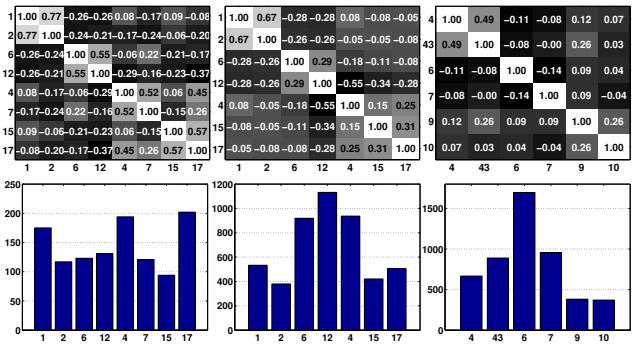
4.1. Datasets

We evaluate the proposed model on three publicly available datasets: Extended Cohn-Kanade (CK+) [13], UNBC-McMaster Shoulder Pain Expression Archive (Shoulder-pain) [15], and Denver Intensity of Spontaneous Facial Actions (DISFA) [18]. These are benchmark datasets of posed (CK+), and spontaneous (Shoulder-pain, DISFA) data, containing a large number of FACS coded AUs. Specifically, CK+ contains 593 video recordings of 123 subjects displaying posed facial expressions in frontal views. The Shoulder-pain dataset 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 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 we treated each AU with intensity larger than zero as active. Fig. 2 depicts the AU relations, and the distribution of the AU activations for the data used from each dataset.

Features: From images in each dataset, 49 fiducial facial points were extracted using the 2D Active Appearance Model [17]. Based on these points, we registered the images to a reference face (average 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 [20] (feature set II) extracted around each facial point from regions of 32×32 pixels. We chose these features as they showed good performance in variety of AU recognition tasks [24]. 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.

Evaluation procedure. We evaluate MC-LVM on a subset of highly correlated AUs, *i.e*., 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), that according to the Prkachin and Solomon formula [15], are associated with pain. We report F1 score as the performance measure. In all our experiments, we performed 5 fold subject independent cross validation.

Models compared. We compare the proposed MC-LVM to GP methods with different learning strategy. Specifically, we compare to the manifold relevance determination (MRD) [6], which uses the variational approximation, and the discriminative shared GP latent variable model (DS-GPVLM) [8] and multi-task latent GP (MT-LGP) [28], which perform exact ML learning. We also compare to the multi-label backpropagation and kNN (k=1), *i.e*. the BPMLL [31] and ML-KNN [32]. Lastly, we compare to the state-of-the-art methods for multiple AU detection:



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

Figure 2. The global AU relations (in terms of correlation coefficients) (upper row), and the distribution of the AU activations within the used datasets (lower row).

hierarchical RBM (HRBM) [29], l_p-regularized multi-task MKL (l_p-MTMKL) [34] and joint patch multi-label learning (JPML) [35]. For the single input methods, we concatenated the two feature sets. For the kernel-based methods, we used the RBF kernel (in l_p-MTMKL we also used the polynomial kernel). Due to the high learning complexity of l_p-MTMKL, we followed the training scheme in [34] where 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 α and λ_C, λ_R parameters, as well as the size of the latent space (set to 8D) were found via a validation procedure.

4.2. Qualitative Results

Fig.3 (left) shows the convergence of the learning criterion in MC-LVM as a function of the used samples during training on the CK+ dataset. We see that for small number of samples, the model does not converge to a minimum. This is expected, since with few samples $(100 - 500)$ the posterior in Eq.(3) cannot be approximated well. By increasing the number of samples to 1000, the model converges, and does not change considerably after that. Thus, we fixed the number of samples to 1000. Fig.3 (right) shows the effect of changing α on the discriminative power of the model, for all three datasets. We observe that the model prefers a weighted conditional distribution, than fully selecting the generative/discriminative component. The optimal value of α is 0.6 for the posed, and 0.8 for the spontaneous data. This difference is because in the case of the naturalistic data (DISFA, Shoulder-pain), the model puts less focus on explaining the unnecessary (for the AU detection) variations (*e.g*., head pose) of the input features. Therefore, the influence of the generative component is lower (higher α) than in the case of the posed expressions from CK+.

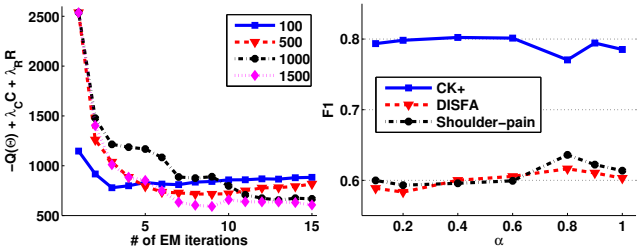


Figure 3. The cost function of the MC-LVM for different number of samples used to estimate the posterior (left), and average F1 score for multiple AU detection as a function of the regularization parameter α (right).

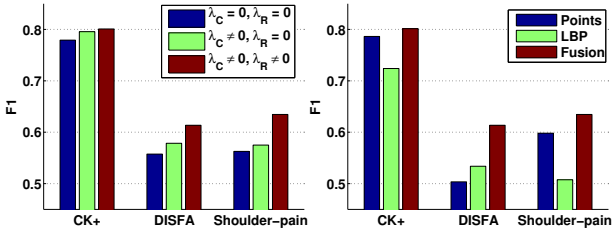


Figure 4. Average F1 score on all three datasets for different settings of the proposed MC-LVM. The effect of the relational constraints (left), and the feature fusion (right) to the joint AU detection task.

Table 1. F1 score for joint AU detection on CK+ dataset.

Methods (I+II)	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**
MC-LVM (SO)	86.06	**88.37**	**82.93**	70.80	57.27	87.16	73.26	85.57	78.93
MRD [6]	80.72	79.18	69.93	69.81	53.24	77.83	65.70	85.20	72.70
MT-LGP [28]	**89.12**	83.70	79.79	67.16	60.89	80.53	64.63	85.97	76.47
DS-GPLVM [8]	87.41	81.78	79.70	68.48	63.29	81.04	60.33	84.29	76.17
HRBM [29]	87.62	84.00	74.10	62.90	50.74	82.38	66.06	84.56	74.04
l_p-MTMKL [34]	87.50	85.50	51.43	**72.65**	58.82	85.95	74.21	75.44	73.93
BPMLL [31]	75.41	84.31	64.85	69.14	**64.34**	83.98	69.50	76.25	73.47
ML-KNN [32]	76.83	84.34	63.28	67.23	53.19	82.88	65.88	78.71	71.54
JPML* [35]	91.2	96.5	-	75.6	50.9	80.4	76.8	80.1	78.8

In Fig.4 (left) we see the effect of the introduced relational constraints on the model's performance, on all three datasets. 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 the topological constraint ($\lambda_C \neq 0$), MC-LVM unravels a better representation of the data in the manifold, which results in higher F1 scores. Finally, with the addition of the global relational constraint ($\lambda_C, \lambda_R \neq 0$) MC-LVM achieves the highest scores. Note that the difference is more pronounced in data from DISFA and Shoulder-pain, which evidences the importance of modeling the global relations for the detection of spontaneous (more subtle) AUs. We continue by evaluating the effectiveness of the proposed MC-LVM on the feature fusion task. To this end, we learn the MC-LVM in single, and multi-input settings. Fig.4 (right) shows the average performance of the model on all three datasets, for the 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 large variations in head pose) in the task of multiple AU detection. This is because, by concatenating the 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 shared latent space with fused information from the global geometrical descriptors and the local patch-related histograms. This results in the highest F1 score, with significant improvement on the spontaneous data of DISFA and Shoulder-pain.

4.3. 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 on the target task. From Table 1, the ML-based methods, *i.e.*, the MT-LGP [28] and DS-GPLVM [8], achieve similar performance on average and per AU, since they are based on the same learning scheme. On the other hand, MRD [6], uses a variational distribution to approximate a manifold shared across multiple inputs and outputs, without any constraints over the latent variables. By contrast, the combination of the approximate learning with the relational constraints used in the proposed MC-LVM results in a significant increase in the performance. We partly attribute this to the explicit modeling of AU co-occurrences through the introduced constraints, as well as the multi-conditional learning using 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 it achieves comparable scores to the multi-output. Finally, the state-of-the-art models for joint AU detection, *i.e.* the HRBM and l_p-MTMKL, improve the detection of specific AUs (AU1,6). Yet, they achieve lower results compared to the proposed MC-LVM. HRBM cannot handle simultaneously the fusion of the *concatenated* features and the modeling of the AU dependencies using binary latent variables. l_p-MTMKL, due to its modeling complexity, it is trained on subsets of AUs (as mentioned above) which affects its ability to capture all AU relations. More importantly, in contrast to MC-LVM, these two models lack the generative component, which, evidently, acts as a powerful regularizer. The results of JPML were obtained from [35], thus, they are not directly comparable to the other models. Yet, we report its performance as a reference to the state-of-the-art. The baseline models, BPMLL and ML-KNN, report the lower 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 for joint AU detection on *all* 17 AUs from CK+ (l_p-MTMKL cannot be evaluated

Table 2. 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**
HRBM [29]	**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 3. F1 score for joint AU detection on DISFA and Shoulder-pain datasets.

Methods (I+II)	DISFA dataset								Shoulder-pain dataset						
	AU1	AU2	AU4	AU6	AU12	AU15	AU17	Avg.	AU4	AU6	AU7	AU9	AU10	AU43	Avg.
MC-LVM	**58.55**	**62.99**	**72.85**	52.32	**84.74**	**49.44**	**48.63**	**61.36**	47.20	**97.75**	67.88	**37.13**	58.23	**72.51**	**63.45**
MC-LVM (SO)	35.50	52.68	70.99	54.67	82.58	37.11	47.76	54.47	**57.76**	95.57	63.59	34.54	49.93	64.49	60.98
MT-LGP [28]	41.44	36.84	61.19	45.98	49.78	40.12	43.01	45.48	50.42	50.48	63.52	33.38	**61.62**	61.00	53.40
HRBM [29]	39.67	55.92	61.56	54.01	79.16	38.72	38.82	52.55	47.20	93.93	63.67	29.80	52.39	69.54	59.42
l_p-MTMKL [34]	42.21	45.81	47.18	**62.79**	76.33	34.47	41.40	50.03	37.69	**97.75**	**70.08**	33.28	41.79	44.03	54.10

on this experiment due to its learning complexity). As we can see in Table 2, modeling of the remaining (less frequent) AUs affects the overall performance of both MC-LVM and HRBM, which suffer a drop of 8.6% and 7.6%, respectively. However, MC-LVM outperforms HRBM on 14 out of 17 AUs, which demonstrates the ability of the former to better model the relations among AUs, even in a larger scale.

4.4. Model Comparisons on Spontaneous Data

We further investigate the models' performance on spontaneous data from DISFA and Shoulder-pain. We focus here on the best performing methods from Table 1. From Table 3, we 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, and demonstrates the difference between posed and spontaneous expressions. We also observe from Figure 2 that the distribution of the activated AUs is more imbalanced than the posed dataset. This imposes an additional challenge since data for certain AUs (*e.g.*, AU2,15 for DISFA, and AU9,10 for Shoulder-pain) are limited compared to others, and thus, the models need to give more emphasis on the AU co-occurrences for their detection. Hence, the single output MC-LVM reports low scores for the aforementioned AUs in both datasets. On the other hand, with limited data the modeling of the global AU relations is even harder task. HRBM is adversely affected by this issue, and it performs close to the single output model. l_p-MTMKL, reports even lower results (especially in Shoulder-pain), due to not modeling global relations. MT-LGP fails to model explicitly the relations between AUs, resulting in low scores. On the other hand, it is evidenced that the proposed MC-LVM is more robust to the data imbalance, and can better discover the AU relations, which in turn gives the best average scores.

5. Discussion and Conclusions

We proposed a novel multi-conditional latent variable model that exploits successfully the non-parametric probabilistic framework of GPs to perform multi-conditional subspace learning for efficient feature fusion and joint AU detection. By assuming conditional independence given the subspace of AUs, MC-LVM allows each feature set to be described via feature-specific 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 of the model to act in concert – the generative component has the key role to unravel the shared subspace of different feature sets, while the discriminative component endows the subspace with the relational information about the output labels. Consequently, this enables MC-LVM to learn the structure of a discriminative subspace that is optimized for multiple AU detection, while being effectively regularized by the generative component. We demonstrated the effectiveness of these properties on three publicly available datasets by showing that the proposed model outperforms the existing works for multiple AU detection, and several methods for feature fusion and multi-label learning. Finally, we showed that the proposed MC-LVM scales well with a large number of AUs, without significant increase in its computational complexity.

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. Also, automatic balancing of the conditional distributions in the model is another direction to pursue. These are going to be the focus of our future work.

Acknowledgment

This work has been funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). The work by S. Eleftheriadis is further supported by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA).

References

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

[2] 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 *IEEE Conf. on CVPR*, volume 2, pages 568–573, 2005.

[3] L. Bo and C. Sminchisescu. Supervised spectral latent variable models. In *Int'l Conf. on AISTATS*, pages 33–40, 2009.

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

[5] F. R. Chung. Spectral graph theory. *American Mathematical Society*, 1997.

[6] A. Damianou, C. H. Ek, M. Titsias, and N. Lawrence. Manifold relevance determination. In *ICML*, pages 145–152, 2012.

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

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

[9] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *SIGKDD*, pages 109–117. ACM, 2004.

[10] X. He, D. Cai, Y. Shao, H. Bao, and J. Han. Laplacian regularized gaussian mixture model for data clustering. *IEEE TKDE*, 23(9):1406–1418, 2011.

[11] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE TPAMI*, 32(11):1940–1954, 2010.

[12] N. D. Lawrence and J. Q. Candela. Local distance preservation in the gp-lvm through back constraints. In *ICML*, volume 148, pages 513–520. ACM, 2006.

[13] 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 *IEEE Conf. on CVPR'W*, pages 94–101, 2010.

[14] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE Trans. on SMCB, Part B*, 41(3):664–674, 2011.

[15] 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 *IEEE Int'l Conf. on AFGR*, pages 57–64, 2011.

[16] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. In *Int'l Conf. on AFGR*, pages 336–342, 2011.

[17] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004.

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

[19] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *Proc. of AAAI*, volume 21, page 433, 2006.

[20] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 24(7):971–987, 2002.

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

[22] O. Rudovic, V. Pavlovic, and M. Pantic. Kernel conditional ordinal random fields for temporal segmentation of facial action units. In *ECCV'W12*, pages 260–269, 2012.

[23] K. R. Scherer and P. Ekman. *Handbook of methods in nonverbal behavior research*, volume 2. Cambridge University Press, 1982.

[24] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *IEEE Trans. on SMCB, Part B*, 42(4):993–1005, 2012.

[25] M. S. Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 2010.

[26] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE TPAMI*, 29(10):1683–1699, 2007.

[27] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *IJDWM*, 3(3):1–13, 2007.

[28] R. Urtasun, A. Quattoni, N. Lawrence, and T. Darrell. Transferring nonlinear representations using gaussian processes with a shared latent space. Technical Report MIT-CSAIL-TR-08-020, 2008.

[29] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *IEEE ICCV*, pages 3304–3311, 2013.

[30] J. Zhang, Z. Ghahramani, and Y. Yang. Flexible latent variable models for multi-task learning. *Machine Learning*, 73(3):221–242, 2008.

[31] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE TKDE*, 18(10):1338–1351, 2006.

[32] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[33] X. Zhang and M. H. Mahoor. Simultaneous detection of multiple facial action units via hierarchical task structure learning. In *IEEE ICPR*, pages 1863–1868, 2014.

[34] X. Zhang, M. H. Mahoor, S. M. Mavadati, and J. F. Cohn. An lp-norm MTMKL framework for simultaneous detection of multiple facial action units. In *IEEE WACV*, pages 1104–1111, 2014.

[35] 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 *IEEE Conf. on CVPR*, pages 2207–2216, 2015.

[36] Y. Zhu, S. Wang, L. Yue, and Q. Ji. Multiple-facial action unit recognition by shared feature learning and semantic relation modeling. In *IEEE ICPR*, pages 1663–1668, 2014.