To appear, *IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2015.*

Exploiting Sparsity and Co-occurrence Structure for Action Unit Recognition

Yale Song [†1], Daniel McDuff [†2], Deepak Vasisht [3], and Ashish Kapoor [4]
[1]Yahoo! Research [2]Affectiva [3]MIT CSAIL [4]Microsoft Research

***Abstract*— We present a novel Bayesian framework for facial action unit recognition. The first key observation behind this work is sparsity: out of possible 45 (and more) facial action units, only very few are active at any moment. The second is the strong statistical co-occurrence structure: most facial expressions are made by common combinations of facial action units, so knowing the presence of one can act as a strong prior for inferring the presence of others. We developed a novel Bayesian graphical model that encodes these two natural aspects of facial action units via compressed sensing and group-wise sparsity inducing priors. One crucial aspect of our approach is the allowance of overlapping group structures, which proves useful in dealing with action units that occur frequently across multiple groups. We derive an efficient inference scheme and show how such sparsity and co-occurrence can be automatically learned from data. Experiments on three standard benchmark datasets show superiority over the state-of-the-art.**

I. Introduction

The Facial Action Coding System (FACS) [9] is the most comprehensive catalogue of unique facial actions that correspond to independent motions of the face. FACS enables the measurement and scoring of facial activity in an objective and quantitative way, and is often used to discriminate between subtle differences in facial motion. However, manual labeling of action units (AUs) is extremely time consuming and requires specific training. It is often infeasible to hand label all or even a subset of AUs. Computer vision hopes to alleviate these challenges via automatic AU recognition [30].

This paper exploits two core properties of facial action units. First, we observe that out of a large number of possible AUs, only a few are observed to be present at any moment. For example, even for complex expressions such as disgust or surprise, less than five AUs are activated (see Figure 1). Such sparsity in action unit space can be very informative for the purpose of AU recognition, as a learning machine can focus all its resources towards recovering the most likely AUs. Further, recent advances in compressed sensing [13], [16] have shown how much computational efficiency such sparsity provides without compromising the quality of the results. Our model incorporates compressed sensing in a Bayesian framework and inherits similar advantages, modeling the sparsity in action unit space in a principled manner.

Another important observation is the existence of strong co-occurrence structure in action units, such as AU1+2 when the eyebrows are raised. Figure 1 shows examples of frequently occurring AU combinations. There is much evidence, both theoretical and empirical, of this type of co-occurrence structure. Perhaps the most well-known are the studies by Ekman [10] who showed certain combinations of AUs often occur together in expressions of pain and emotion. Further, our analysis on a spontaneous expression dataset [17] revealed that 10 AU groups occurred in 48% of the time in which more than one AU was present, and in more than 55% of instances of AU7 (lid tightener), it co-occurred with AU4 (corrugator), suggesting only a few groups account for many of the possible AU combinations.

Fig. 1. Facial action units have strong sparsity and co-occurrence structure. Even for complex expressions such as disgust or surprise, less than five out of 45 action units are activated. Further, groups of action units tend to co-occur in similar expressions. We exploit these two properties for facial action unit recognition.

The existence of strong co-occurrence structure suggests that the presence of one AU can act as a strong prior for the presence of others, e.g., detecting AU6 (cheek raiser) is difficult due to very subtle changes in appearance and geometry of the face, but it is known to co-occur quite frequently with AU12 (in a "Duchenne smile"), so the presence of AU12 increases the chance of AU6 being activated. Co-occurrence information has recently started to prove useful in sparsity-based methods [14], where groups of variables are constrained to be zero/non-zero simultaneously. This has a natural connection to AU recognition: we want frequently co-occurring AUs to be active simultaneously. One crucial requirement is a way to deal with overlapping groups, as an example AU25 (lips apart) appears in many different combinations. Our proposed method leverages the co-occurrence structure in action unit space, and naturally handles overlapping AU groups (detailed in Section III-D).

We developed a Bayesian framework that simultaneously handles the properties of sparsity and co-occurrence structure

[†] indicates equal contribution. This work was done when Y. Song and D. McDuff were at MIT CSAIL and MIT Media Lab, respectively.

in a principled manner, using compressed sensing and group-wise sparsity inducing priors. We extend the recent work in Bayesian Compressed Sensing [16] by incorporating a multivariate Normal-Gamma hierarchical prior term, and show that the previous work [16] is a special case of our model. Finding the true underlying group structure is an open problem in the group sparsity literature and many resort to a manual definition [15]. Instead, we automatically learn the optimal group definitions using the co-occurrence statistics computed from an independent, large-scale dataset of spontaneous expressions [17], and show empirically that it generalizes well across datasets. Additionally, our model can handle partially labeled data, potentially reducing the labeling burden on FACS coders. Also, the uncertainties are maintained over the course of the Bayesian inference; thus, information from (a) the observations, (b) compressed AU labels, and (c) group sparsity constraints are combined in a principled manner. To the best of our knowledge, this work is the first to exploit both the sparsity and co-occurrence structure of AUs. In summary, our main contributions are:

- A Bayesian model that exploits sparsity and co-occurrence structure for detecting AUs, using compressed sensing and group-wise sparsity inducing priors.
- An optimal AU group structure automatically learned from co-occurrence statistics of independent data.
- Superior performance over the state-of-the-art on the CK+ [19], G.-FERA [3], and DISFA [21] datasets.
- MATLAB code available at http://people.csail.mit.edu/yalesong/fg15

II. Related Work

A comprehensive review on facial expression recognition can be found in [30]. Most approaches are direct applications of existing classification techniques, such as SVMs [4], [28] and Bayes Nets [26], operating on geometric or appearance features such as histograms of oriented gradients (HOG) and Gabor energy filters. Previous work on AU detection from video includes Valstar and Pantic [27], who demonstrate high agreement with human coders on 15 AUs, and Bartlett *et al.* [4] who use a framework combining Gabor features and SVMs to detect 17 AUs. Valstar *et al.* [29] have also presented a hybrid SVM-HMM system using Gabor features to detect 23 AUs and, in follow-up work [28], included comparisons over 12 AUs. Although sparsity in the feature space has been addressed in facial expression analysis [20], we would like to point out that, unlike our approach, these methods neither model the sparsity of AU space, nor encode the co-occurrence structure.

Related to the task of modeling AU co-occurrence statistics is multi-task learning [7]. Tian *et al.* [24] is perhaps the most direct application of concepts in multitask learning to AU recognition, where a single Neural Network with multiple outputs was trained. We'd like to highlight that such methods mostly provide a boost in accuracy via shared representation as opposed to direct encoding of the co-occurrence property. Tong *et al.* [26], [25] presented a dynamic Bayesian network (DBN) for inference and showed that learning the relationship between AUs strengthens prediction. Li *et al.* [18] extended the DBN approach for measuring the intensity of action units. Missing from such methods is the capability to address sparsity, which we show to be very useful for AU recognition.

Modeling sparsity in the label space has only been addressed recently. Compressed sensing is perhaps one of the more promising methods [13], [1], [16]. Hsu *et al.* [13] proposed compressing the sparse label space in order to reduce the multiclass problem into simpler regression tasks. Our work builds upon this line of research and extends the Bayesian framework proposed by Kapoor *et al.* [16]. The key differentiating aspect of our work is to explicitly model and learn the co-occurrence structure, which was missing from the earlier work. To the best of our knowledge, this paper is the first to propose exploiting both the sparsity and co-occurrence structure of AUs. Finally, the ability to marginalize over unknown labels allows us to learn good recognition models even with partially observed labels.

The main novelty in this work is the use of group sparsity over facial action units to exploit both the sparsity and co-occurrence structure. Due to its ability to encode group structure in the variables of interest, group sparsity has recently gained much interest [15], [23], [12]. In a Bayesian framework, Raman *et al.* [23] used group-lasso to exploit co-occurring patterns of marker proteins sampled from patients diagnosed with breast cancer. The key difference in our work is the use of regression functions to jointly optimize sparsity and the input-output compatibility. Zhong *et al.* [31] used group sparsity over the image space for AU recognition, where the face image is divided into non-overlapping patches and grouped by their conceptual roles in making expressions. Different from their approach, our notion of group sparsity is focused on the output (action unit) space, as opposed to the input (image) space, which enables more direct control of sparsity in action units.

III. Our Approach

We cast the problem of detecting facial action units as a multi-label binary classification problem. In particular, we build upon Bayesian Compressed Sensing (BCS) [16] and extend it to exploit both sparsity and co-occurrence structure via group-wise sparsity inducing priors. Our main technical contribution is the incorporation of multivariate Normal-Gamma hierarchical priors over the output variables to encourage sparsity among overlapping groups of AUs.

We first briefly review the BCS approach, which becomes the foundation of ours (Section III-A). Our proposed Bayesian Group-sparse Compressed Sensing (BGCS) is described next (Section III-B), followed by parameter estimation using variational Bayes (Section III-C). Note that some AUs may appear in multiple different groups (e.g., AU25 in Figure 1); we describe how our model deals with overlapping groups (Section III-D). Finally, we discuss the case with partially-observed AU labels (Section III-E).

Notation: We denote by $\mathbf{x} = [x^1, \cdots, x^d] \in \mathbb{R}^d$ the input data and by $\mathbf{y} = [y^1, \cdots, y^l] \in \{0, 1\}^l$ the corre-

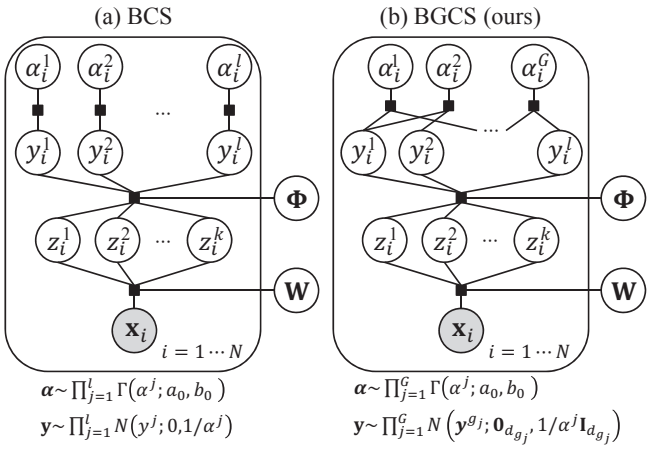

Fig. 2. Factor graph representations of (a) Bayesian Compressed Sensing (BCS) and (b) Bayesian Group-sparse Compressed Sensing (BGCS). The input is $\mathbf{x}$, the multi-label output is $\mathbf{y}$ (fully/partially observed during training, completely unobserved during testing). The sparse label $\mathbf{y}$ is compressed through a sensing matrix $\mathbf{\Phi}$, resulting in the latent variable $\mathbf{z} \approx \mathbf{\Phi y}$. The sparsity is induced via $\boldsymbol{\alpha}$ that follows independent Gamma distributions. The input-output compatibility is learned through a set of regression functions with weight $\mathbf{W}$, mapping the input $\mathbf{x}$ to the compressed output $\mathbf{z}$, i.e., $\mathbf{z} \approx \mathbf{Wx}$. Notice that our BGCS model encourages group-wise sparsity over the output $\mathbf{y}$, allowing us to exploit both sparsity and co-occurrence structure of action units in a principled manner.

sponding multi-output labels, e.g., the presence/absence of AUs. The notation y^j refers to the j-th element of a vector. The sets of input data and output labels are denoted by $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \in \mathbb{R}^{d \times N}$ and $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\} \in \{0,1\}^{l \times N}$, respectively. Further, we use subscripts $\mathcal{L}$ and $\mathcal{U}$ to denote labeled and unlabeled data, respectively; $\mathbf{Y} = \mathbf{Y}_{\mathcal{L}} \cup \mathbf{Y}_{\mathcal{U}}$. For clarity of the presentation, we omit the sample index subscripts whenever it is clear from the context.

A. Bayesian Compressed Sensing: A Review

BCS learns, in a Bayesian framework, an input-output mapping function $\mathbf{y} = F(\mathbf{x})$ by considering the following two tasks simultaneously. One is *compressed sensing*: recovering the output signal $\mathbf{y}$, which is assumed to be sparse, from a lower dimensional "compressed" signal $\mathbf{z} \in \mathbb{R}^k$, obtained through a "sensing" matrix $\mathbf{\Phi} \in \mathbb{R}^{k \times l}$ that satisfies the restricted isometry property [6], $\mathbf{z} \approx \mathbf{\Phi y}$. Another is *regression*: learning the relationship between the input $\mathbf{x}$ and the compressed signal $\mathbf{z}$ using a set of k regression functions with weights $\mathbf{W} \in \mathbb{R}^{k \times d}$, $\mathbf{z} \approx \mathbf{Wx}$.

Figure 2 (a) illustrates the BCS model; the upper part (involving $\boldsymbol{\alpha}, \mathbf{y}, \mathbf{z}, \mathbf{\Phi}$) corresponds to compressed sensing, the lower part (involving $\mathbf{z}, \mathbf{x}, \mathbf{W}$) corresponds to regression. The latent variable $\mathbf{z}$ plays the role of balancing between the two tasks, optimizing the compatibility between the input $\mathbf{x}$ and the compressed output $\mathbf{z}$, while simultaneously encouraging sparsity of the output $\mathbf{y}$, i.e., $\mathbf{\Phi y} \approx \mathbf{z} \approx \mathbf{Wx}$. The joint Bayesian formulation makes the BCS capable of capturing important statistical relationships amongst different variables of interest, improving accuracy [16].

To induce sparsity, BCS defines a zero-mean univariate Gaussian prior over each element of the output, $y^j \sim \mathcal{N}(0, 1/\alpha^j)$, where the precision (inverse variance) α^j follows the Gamma distribution $\alpha^j \sim \Gamma(a_0, b_0)$. Each Normal-Gamma prior over y^j is assumed to be independent of each other, and the prior over $\mathbf{y}$ has the form:

$$p(\mathbf{y}) = \prod_{j=1}^{l} p(y^j) = \prod_{j=1}^{l} \int_0^\infty p(y^j|\alpha^j) p(\alpha^j) d\alpha^j \quad (1)$$

$$= \prod_{j=1}^{l} \int_0^\infty \mathcal{N}(y^j; 0, 1/\alpha^j) \Gamma(\alpha^j; a_0, b_0) d\alpha^j \quad (2)$$

This is also known as the Gaussian scale mixture, where the mixing distribution on the precision $p(\alpha^j)$ is the Gamma distribution [22]. The integral form in Equation 2 follows the student-t distribution that has a significant probability mass around zero. Consequently, with a proper choice of a_0 and b_0, most elements of $\mathbf{y}$ are going to be zero unless otherwise necessary to describe the observed data, encouraging sparsity of the output. Notice that sparsity is encouraged only element-wise due to the independence assumption among different elements of $\mathbf{y}$, limiting its use for exploiting co-occurrence structure among the output elements.

B. Bayesian Group-Sparse Compressed Sensing

To exploit both sparsity and co-occurrence structure, our Bayesian Group-sparse Compressed Sensing (BGCS) alleviates the element-wise sparsity assumption and instead defines a group-wise sparsity prior over the output space.

Let $\mathcal{G} = \{g_1, \cdots, g_G\}$ be a set of groups, with each group g having a set of d_g indices. We define a zero-mean d_g-dimensional multivariate Gaussian prior over each group of the output, $\mathbf{y}^g \sim \mathcal{N}(\mathbf{0}_{d_g}, \alpha^{-1} \mathbf{I}_{d_g})$, where $\mathbf{y}^g$ is a sub-vector of $\mathbf{y}$ formed by taking g elements, $\mathbf{0}_{d_g}$ is a zero-vector of length d_g, and $\mathbf{I}_{d_g}$ is the $d_g \times d_g$ identity matrix. The precision parameter α is again assumed to follow a Gamma distribution, $\alpha \sim \Gamma(a_0, b_0)$. Assuming independence between groups, we express the prior over $\mathbf{y}$ as:

$$p(\mathbf{y}) = \prod_{j=1}^{G} p(\mathbf{y}^{g_j}) = \prod_{j=1}^{G} \int_0^\infty p(\mathbf{y}^{g_j}|\alpha^j) p(\alpha^j) d\alpha^j \quad (3)$$

$$= \prod_{j=1}^{G} \int_0^\infty \mathcal{N}(\mathbf{y}^{g_j}; \mathbf{0}_{d_{g_j}}, \frac{1}{\alpha^j} \mathbf{I}_{d_{g_j}}) \Gamma(\alpha^j; a_0, b_0) d\alpha^j \quad (4)$$

Notice that elements within a group $\mathbf{y}^g$ are not independent anymore; rather, as can be seen from the integral form in Equation 4, they follow the multivariate student-t distribution. Consequently, elements within the same group will tend to zero simultaneously, encouraging group-wise sparsity; this is the desirable property to achieve exploiting both sparsity and co-occurrence structure of action units. Also notice that, when $\sum_j d_{g_j} = l$ for $\forall j,\ d_{g_j} = 1$, our model reduces to the conventional BCS model, ignoring any structural information among action units, making the BCS a special case of ours.

Figure 2 (b) shows the factor graph representation of our BGCS model. The lower part remains identical to that of the BCS; we are still optimizing the compatibility between the input $\mathbf{x}$ and the compressed output $\mathbf{z}$. Different from the

BCS, however, sparsity is induced for each group of elements $\mathbf{y}^g$ with a multivariate Normal-Gamma distribution.

We now formalize our BGCS model. The observables include input data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and the sensing matrix $\mathbf{\Phi}$; in this work, $\mathbf{\Phi}$ is set with random values between -1 and 1, which satisfies the restricted isometry property [6]. The unknowns include output labels $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$, compressed labels $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^N$, precision parameters $\mathbf{A} = \{\boldsymbol{\alpha}_i\}_{i=1}^N$, and the regression weight $\mathbf{W}$. With these variables, the posterior $p(\cdot) = p(\mathbf{Y}, \mathbf{Z}, \mathbf{A}, \mathbf{W}|\mathbf{X}, \mathbf{\Phi})$ is expressed as:

$$p(\cdot) = \frac{1}{\mathcal{Z}} p(\mathbf{W}) \prod_{i=1}^{N} f_{\mathbf{x}_i}(\mathbf{W}, \mathbf{z}_i) g_{\mathbf{\Phi}}(\mathbf{y}_i, \mathbf{z}_i) p(\mathbf{y}_i) \quad (5)$$

where $\mathcal{Z}$ is a normalization term. The prior over the regression weight $\mathbf{W}$ is defined for each of the k regression functions as the d-dimensional spherical Gaussian distribution, $p(\mathbf{W}) = \prod_{j=1}^{k} \mathcal{N}(\mathbf{w}^j; \mathbf{0}_d, \mathbf{I}_d)$. The two potential functions $f_{\mathbf{x}_i}(\mathbf{W}, \mathbf{z}_i)$ and $g_{\mathbf{\Phi}}(\mathbf{y}_i, \mathbf{z}_i)$ are defined as:

$$f_{\mathbf{x}_i}(\mathbf{W}, \mathbf{z}_i) = e^{-\frac{\|\mathbf{W}\mathbf{x}_i - \mathbf{z}_i\|^2}{2\sigma^2}}, g_{\mathbf{\Phi}}(\mathbf{y}_i, \mathbf{z}_i) = e^{-\frac{\|\mathbf{\Phi}\mathbf{y}_i - \mathbf{z}_i\|^2}{2\chi^2}} \quad (6)$$

Intuitively, $f_{\mathbf{x}_i}(\mathbf{W}, \mathbf{z}_i)$ measures the compatibility between the input $\mathbf{x}$ and the compressed output $\mathbf{z}$ determined by $\mathbf{W}$, while $g_{\mathbf{\Phi}}(\mathbf{y}_i, \mathbf{z}_i)$ measures the compatibility between $\mathbf{y}$ and $\mathbf{z}$ compressed by $\mathbf{\Phi}$. The variance terms σ^2 and χ^2 control how tight we want each compatibility to be; by changing the variance terms we can fine-tune the relative importance of the two potentials. We find the optimal values of the two terms via cross-validation.

Finally, the group-sparse prior over $\mathbf{y}_i$ is given in Equation 4. The two terms a_0 and b_0 of the prior $p(\mathbf{y}_i)$ control the shape and (inverse) scale of the Gamma distribution, respectively, determining the level of sparsity over the groups. Following [16], we initialize values of the two terms to 10^{-6}, which makes α close to a diffuse (non-informative) prior (i.e., a distribution of the parameter with equal probability for each possible value), then optimize them via Bayesian inference, described below.

C. Variational Bayes Inference

Given input data $\mathbf{X}$ with observed and unknown labels $\mathbf{Y}_\mathcal{L}$ and $\mathbf{Y}_\mathcal{U}$, respectively, the goal of the inference is to compute the posterior over the unlabeled data $p(\mathbf{Y}_\mathcal{U}|\mathbf{X}, \mathbf{Y}_\mathcal{L})$ by integrating out all other latent variables of the model. In general, performing an exact inference is intractable for forms that involve the product of Gaussian and Gamma distributions [22]; thus, approximate methods are commonly used. In this work, we perform approximate inference, maximizing the variational lower bound by making a fully factorized (i.e., mean field) approximation of the posterior. This method is commonly called the Variational Bayes (VB).

Let $\boldsymbol{\xi} = \{\mathbf{Y}_\mathcal{U}, \mathbf{Z}, \mathbf{A}, \mathbf{W}\}$ be all the unknowns of our model, and $q(\cdot)$ be an approximation of the true posterior $p(\cdot)$. We want to maximize the lower bound $\mathcal{J}(q)$:

$$\mathcal{J}(q) = \int_{\boldsymbol{\xi}} q(\boldsymbol{\xi}) \log \frac{p(\boldsymbol{\xi}|\mathbf{X}, \mathbf{\Phi})}{q(\boldsymbol{\xi})} \leq \log \int_{\boldsymbol{\xi}} p(\boldsymbol{\xi}|\mathbf{X}, \mathbf{\Phi}) \quad (7)$$

The mean field approximation $q(\boldsymbol{\xi})$ of the true posterior has the following fully factorized form:

$$q(\boldsymbol{\xi}) = q(\mathbf{Y}_\mathcal{U}) q(\mathbf{Z}) q(\mathbf{A}) q(\mathbf{W}) \quad (8)$$

where further factorizations are made per-data for $q(\mathbf{Y}_\mathcal{U}) = \prod_{i \in \mathcal{U}} q(\mathbf{y}_i)$ and $q(\mathbf{Z}) = \prod_{i \in \mathcal{L} \cup \mathcal{U}} q(\mathbf{z}_i)$; per-data and per-group for $q(\mathbf{A}) = \prod_{i \in \mathcal{L} \cup \mathcal{U}} \prod_{j=1}^{G} q(\alpha_i^j)$; and per-function for $q(\mathbf{W}) = \prod_{j=1}^{k} q(\mathbf{w}^j)$.

VB optimizes the objective by iteratively updating each of the factorized distributions $q(\cdot)$. Specifically, at each iteration t, the update rules for the Gaussian terms $q(\mathbf{y}_i) = \mathcal{N}(\mu_{\mathbf{y}_i}, \Sigma_{\mathbf{y}_i})$ (and similarly $q(\mathbf{z}_i)$ and $q(\mathbf{w}^j)$) and the Gamma term $q(\alpha_i^j) = \Gamma(a_{ij}, b_{ij})$ is:

Update $q^{t+1}(\mathbf{y}_i)$:

$$\Sigma_{\mathbf{y}_i}^{t+1} = \left[\text{diag}(\mathbb{E}[\overline{\boldsymbol{\alpha}}_i^t]) + \chi^{-2}\mathbf{\Phi}^T\mathbf{\Phi}\right]^{-1}$$
$$\mu_{\mathbf{y}_i}^{t+1} = \Sigma_{\mathbf{y}_i}^{t+1}\chi^{-2}\mathbf{\Phi}^T\mu_{\mathbf{z}_i}^t$$

Update $q^{t+1}(\mathbf{z}_i)$:

$$\Sigma_{\mathbf{z}_i}^{t+1} = \left[\sigma^{-2}\mathbf{I}_k + \chi^{-2}\mathbf{I}_k\right]^{-1}$$
$$\mu_{\mathbf{z}_i}^{t+1} = \Sigma_{\mathbf{z}_i}^{t+1}\left[\sigma^{-2}\mu_{\mathbf{W}}^t\mathbf{x}_i + \chi^{-2}\mathbf{\Phi}\mu_{\mathbf{y}_i}^{t+1}\right]$$

Update $q^{t+1}(\mathbf{w}^j)$:

$$\Sigma_{\mathbf{w}_i}^{t+1} = \left[\sigma^{-2}\mathbf{X}\mathbf{X}^T + \mathbf{I}_d\right]^{-1}$$
$$\mu_{\mathbf{w}^j}^{t+1} = \Sigma_{\mathbf{w}^j}^{t+1}\sigma^{-2}\mathbf{X}[\mu_{\mathbf{Z}}^{t+1}(j,:)]^T$$

Update $q^{t+1}(\alpha_i^j)$:

$$a_{ij}^{t+1} = a_{ij}^0 + \frac{1}{2}d_{g_j}$$
$$b_{ij}^{t+1} = b_{ij}^0 + \frac{1}{2}\left[\|\mu_{\mathbf{y}_i^j}^{t+1}\|_2^2 + \text{tr}\left(\Sigma_{\mathbf{y}_i^j}^{t+1}(g_j, g_j)\right)\right]$$

where the vector $\overline{\boldsymbol{\alpha}}_i$ of length l is formed by repeating each α_i^j d_{g_j}-times. Note that the above update rules assume that the groups do not overlap; below we describe how to deal with overlapping group structure. The most expensive step in this scheme is the inversion of a $d \times d$ matrix for updating $\Sigma_{\mathbf{w}_i}^{t+1}$; this is an $O(d^3)$ update that is independent of the number of labels. The inversion of a $l \times l$ matrix for updating $\Sigma_{\mathbf{y}_i}^{t+1}$ is not needed when the labels are fully observed, e.g., in training. This inference scheme, together with compressed sensing, makes our model particularly efficient in dealing with a high-dimensional output space.

Alternating between the above updates can be seen as message passing between different layers of the factor graph shown in Figure 2 (b). The core idea is to determine a configuration of latent variables that fuses information for both the feature space and the label space that is group-wise sparse. Specifically, the latent variables $\mathbf{Z}$ are constrained by feature vectors $\mathbf{X}$ via the linear regression functions $\mathbf{W}$. These latent variables also need to align themselves with the output labels $\mathbf{Y}$ through the sensing matrix $\mathbf{\Phi}$ and the group-wise sparsity by $\boldsymbol{\alpha}$. Consequently, the resulting inference procedure over the graphical model leads to a labeling of

AUs that captures our beliefs about the sparsity and co-occurrence structure of the facial action units.

D. Overlapping Groups

The assumption that AU groups do not overlap may pose a serious problem in AU recognition because some AUs often appear in multiple groups. For example, AU4 (eye brow lowerer) could co-occur with AU7 (lid tightener) and/or AU45 (blink) in three different settings, i.e., AU 4+7, AU 4+45, and AU 4+7+45. However, the strict disjoint-group assumption would allow only one of the three AU groups to exist. Therefore, in order to encode co-occurrence structure correctly, we must allow overlapping group definitions.

Similar to Jacob *et al.* [15], we handle overlapping groups by explicitly duplicating the label vector $\mathbf{y}$ and the sensing matrix $\mathbf{\Phi}$ that correspond to the elements belonging to multiple groups. Specifically, we define $\mathbf{y}' = [\mathbf{y}^{g_1}; \cdots; \mathbf{y}^{g_G}] \in \mathbb{R}^{d_G}$ and $\mathbf{\Phi}' = [\mathbf{\Phi}(:, g_1); \cdots; \mathbf{\Phi}(:, g_G)] \in \mathbb{R}^{k \times d_G}$, where $d_G = \sum_j d_j$. With this modification, the inference procedure of Section III-C remains the same, except for the update rules of duplicated output variables $q(\mathbf{y}')$:

$$\Sigma_{\mathbf{y}'_i}^{t+1} = \left[\mathbb{E}[\overline{\boldsymbol{\alpha}}_i^t]\mathbf{I}_{d_G} + \chi^{-2}\mathbf{\Phi}'^T\mathbf{\Phi}'\right]^{-1}$$
$$\mu_{\mathbf{y}'_i}^{t+1} = \Sigma_{\mathbf{y}'_i}^{t+1}\chi^{-2}\mathbf{\Phi}'^T\mu_{\mathbf{z}_i}^t$$

We then compute $\Sigma_{\mathbf{y}_i}^{t+1}$ and $\mu_{\mathbf{y}_i}^{t+1}$ by marginalizing over the duplicated elements of $\mathbf{y}'$. The duplication method is simple to implement and works well for a small number of overlapping groups [15]; when many groups overlap, however, other methods such as marginalizing prior inverse variances [2] are used for better scalability.

Our method of handling overlapping groups has an important property: it allows AUs in the same group to have different prior distributions, as it should be, not an identical one. To see this, consider a set of groups that jointly contains certain AUs that occur more frequently (e.g., AU25). Because we compute $\mu_{\mathbf{y}_i}^{t+1}$ by marginalizing over the duplicated elements of all overlapping groups, AUs that appear across multiple groups end up having a higher posterior probability.

E. Handling Partially Labeled Data

Our approach naturally handles partially observed labels $\mathbf{Y}_{\mathcal{U}}$ by marginalizing over the unobserved values as a part of the inference procedure. Consider an input $\mathbf{x}_i$ with observed labels $\mathbf{y}_i^o$ and unobserved labels $\mathbf{y}_i^u$. Then, all the above mentioned update steps remain the same except for the update equation of $\mu_{\mathbf{z}_i}^{t+1}$, which now becomes:

$$\mu_{\mathbf{i}_p}^{t+1} = \Sigma_{\mathbf{i}_p}^{t+1}\left[\sigma^{-2}\mu_{\mathbf{W}}^t\mathbf{x}_i + \chi^{-2}\mathbf{\Phi}^{uo}[\mu_{\mathbf{y}_i^u}^{t+1}; \mathbf{y}_i^o]\right]$$

where $\mathbf{\Phi}^{uo}$ represents a reordering of the sensing matrix $\mathbf{\Phi}$ as per the indices of the unobserved and observed labels.

IV. Obtaining Groups of Action Units

One way to obtain AU groups is to use existing definitions of prototypical expressions of emotion from the psychology literature [11]. However, these descriptions may miss naturally occurring combinations that do not make up a prototypic expression of emotion. Instead, we obtain AU groups by computing co-occurrence statistics from an independent, large-scale dataset of spontaneous facial expressions [17]. By definition, the co-occurrence statistics capture both the commonly occurring combinations (present groups) and those do not occur together (absent groups). Thus, our approach effectively models both the co-occurrence and mutually exclusive relationships among action units.

TABLE I

Definitions of the 24 FACS labels considered in this paper.

AU	Definition	AU	Definition	AU	Definition
1	Inner brow raiser	11	Nasolabial deepen.	22	Lip funneler
2	Outer brow raiser	12	Lip corner puller	23	Lip tightener
4	Brow lowerer	14	Dimpler	24	Lip pressor
5	Upper lid raiser	15	Lip corner depress.	25	Lips part
6	Cheek raiser	16	Lower lip depress.	26	Jaw drop
7	Lid tightener	17	Chin Raiser	27	Mouth stretch
9	Nose wrinkler	18	Lip puckerer	43	Eyes closed
10	Upper lip raiser	20	Lip stretcher	45	Blink

We used an independent dataset provided by Kassam [17], a dataset of facial expressions labeled by two certified FACS coders. It contains video recordings of subjects watching emotion eliciting movie clips (704 videos; 88 subjects times 8 clips), with a total length of 61,816 seconds. Frames were FACS coded for 65 AUs at one second intervals; the coders had to agree on the labels. This yielded a total of 61,816 label instances.

Considering 24 AUs (see Table I) and excluding 30,134 instances with no AU activation, about half the rest (15,420 instances) contained more than one active AU, showing strong AU co-occurrence structure in spontaneous facial expressions. The eight most common AU groups were: AU25,26 (1,782 instances); AU4,7 (1,421); AU4,45 (1,207); AU1,2 (796); AU12,45 (587); AU12,25 (538); AU6,7,12,25,26 (518); and AU4,7,45 (411).

To obtain AU groups $\mathcal{G}$, we used *AU-conditional* thresholding, a more robust approach to the class imbalance problem than *joint* thresholding. We computed normalized co-occurrence statistics conditioned on each AU j, $p(\text{AUs}|j)$. We used a threshold parameter θ to rule out those AUs that co-occurred less than θ percentage of the time conditioned on j, i.e., $g_j = \arg p(\text{AU}|j) \geq \theta$. The approach is different to *joint* thresholding, where all AU groups are considered jointly; the resulting AU groups will be dominated by more frequently occurring AUs, failing to capture less common groups. The optimal θ was obtained via cross-validation.

V. Experiments

A. Datasets and Methodology

Our framework is summarized in Figure 3. We used the Nevenvision facial landmark detector[1] to identify 22 facial landmarks within each frame of the video. The face was segmented using the landmarks in rigid locations; an affine warp was performed on the bounded face region; and the segmented face patch was rescaled to 120x120 pixels and

[1]Licensed from Google, Inc.

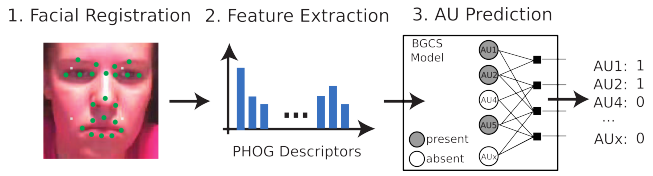

Fig. 3. Our AU recognition framework: (1) the face is registered using 22 automatically detected landmarks; (2) appearance descriptors (PHOG) are extracted; (3) our Bayesian model detects active AUs, exploiting sparsity and co-occurrence structure.

converted to grayscale. We then computed Pyramid Histogram of Gradients (PHOG) [5] features with eight bins on three different pyramid levels from the normalized images.

We used the following datasets in our experiments: **CK+** [19]: The extended Cohn-Kanade (CK+) dataset contains 593 recordings (123 subjects) of posed and non-posed sequences recorded under controlled lighting. We took the last frame (peak expression) from each sequence as these have been FACS coded; this results in 593 frames.

G.-FERA [3]: The GEMEP corpus consists of acted emotion sequences that involve speaking and rigid head motion, which makes it more challenging than the CK+ dataset. We followed the protocol used in the FERA challenge [28], using 87 sequences (5,172 frames; 7 subjects) that were FACS coded and available as training data.

DISFA [21]: The DISFA corpus consists of spontaneous and naturalistic sequences of facial responses to YouTube videos. These sequences are challenging as they tend to be more subtle than acted expressions. We use 27 recordings (130,815 frames; 27 subjects) of spontaneous sequences.

TABLE II

COMPARISON OF MODELS TESTED IN OUR EXPERIMENTS.

Property	SVM	RLS	BCS	BGCS
Sparsity	×	×	✓	✓
Co-occurrence	×	×	×	✓

In addition to evaluating our BGCS model, we selected three baselines to test individual properties in our model; Table II summarizes the different properties.

SVM: We used a linear SVM (one-vs-all) with an option to output probability estimates. The SVM cost term C was cross-validated from the set $C = 10^n, n = [-2 : 1]$. A decision function was defined with a probability threshold δ, cross-validated from the set $\delta = [0 : .05 : 1]$.

RLS: For an approach without the sparsity and co-occurrence properties, we used the regularized least squares (RLS), $\frac{1}{2}\|\mathbf{Y} - \mathbf{WX}\|_F^2 + \lambda\|\mathbf{W}\|_F$, where $\mathbf{W} \in \mathbb{R}^{l\times d}$ is a weight matrix and $\|\cdot\|_F$ is the Frobenius norm. The parameter λ was cross-validated from the set $\lambda = [0 : .1 : 1]$.

BCS: For an approach without the co-occurrence property, we used the BCS [16]. We varied the two scale terms $\chi = 10^n, \sigma = 10^n, n = [-2 : 0]$ (see Equation 6). The compression level was varied among $c = [.2 : .2 : 4]$, which determined the dimension of the compressed space by $k = cd$ with d-dimensional input. Optimal values of all hyper-parameters were determined via cross-validation.

TABLE III

COMPARISON TO THE STATE-OF-THE-ART, ON DIFFERENT SUBSETS OF AUS (SEE THE TEXT FOR THE LIST OF AUS IN EACH SUBSET).

Across subsets of AUs		SVM [8]	MCF [8]	BGCS
CK+	F1 Score	0.71	0.76	**0.90**
(10 AUs)	Accuracy	n/a	n/a	**94.7**
G.-FERA	F1 Score	**0.58**	0.57	0.56
(12 AUs)	Accuracy	n/a	n/a	**76.4**
Across subsets of AUs		AdaBoost [26]	DBN [26]	BGCS
CK+	F1 Score	n/a	n/a	**0.86**
(14 AUs)	Accuracy	91.2	93.3	**93.4**
Across subsets of AUs		SVM [21]		BGCS
DIFSA	F1 Score	n/a		**0.60**
(12 AUs)	Accuracy	85.7		**86.8**

TABLE IV

MEANS AND STANDARD DEVIATIONS COMPARISON USING ALL 24 AUS.

Across all 24 AUs		SVM	RLS	BCS	BGCS
CK+	F1 Score	0.50 (0.14)	0.57 (0.18)	0.63 (0.20)	**0.66 (0.18)**
	Accuracy	85.1 (0.04)	88.2 (0.06)	90.3 (0.05)	**90.5 (0.05)**
G.-FERA	F1 Score	0.39 (0.05)	**0.45 (0.06)**	0.43 (0.07)	0.43 (0.07)
	Accuracy	81.5 (0.03)	82.7 (0.02)	82.8 (0.01)	**83.2 (0.01)**

BGCS (our model): A generalization of BCS with group sparsity, defined with one additional parameter that determines the group structure, the AU-conditional thresholding parameter θ (see Section IV). We cross-validated this from the set $\theta = [.2 : .2 : 1]$. For a fair comparison, other parameters $(\chi, \sigma, c, a_0.b_0)$ were varied as with the BCS.

Note that, except for the SVM, the prediction $\mathbf{Y}^* \in \mathbb{R}^{l\times N}$ includes real-valued regression coefficients, which can be used not only in AU classification but also in AU intensity estimation; this work focuses on classification. We define a decision function $V : \mathbb{R} \to \{0, 1\}$, $V(y; \delta) = 1$ if $y \geq \delta$ and zero otherwise. The parameter δ was cross-validated from the set $\delta = [0 : .05 : 2]$. We performed leave-one-subject-out cross-validation, with data from two subjects for validation and test, respectively, and the rest for training.

B. Results and Discussion

Comparison to state-of-the-art: We compare our model to recent state-of-the-art approaches [8], [26], [21]. In these, different numbers of AUs were considered; for fair comparison we consider the same set of AUs – in [8] 10 AUs {1,2,4,6,7,12,15,17,25,26} for CK+ and 12 AUs {1,2,4,6,7,10,12,15,17,18,25,26} for G.-FERA were used; in [26] 14 AUs {1,2,4,5,6,7,9,12,15,17,23,24,25,27} for CK+ were used; and in [21] 12 AUs {1,2,4,5,6,9,12,15,17,20,25,26} for DISFA were used. Table III shows a comparison of the performances; our BGCS model outperforms all other baselines except the F1 score on G.-FERA dataset.

Evaluation on 24 AUs: We performed classification on 24 AUs (see Table I). Note that this result is rarely reported in the literature, mostly because some AUs are hard to detect, e.g., AU22 (lip funneler) and AU23 (lip tightener). Table IV shows the average F1-scores and accuracies on two of the datasets that have labels for 24 AUs. Our BGCS again tops among the contenders, except for F1-score on G.-FERA. The standard deviations for F1-scores are quite high because

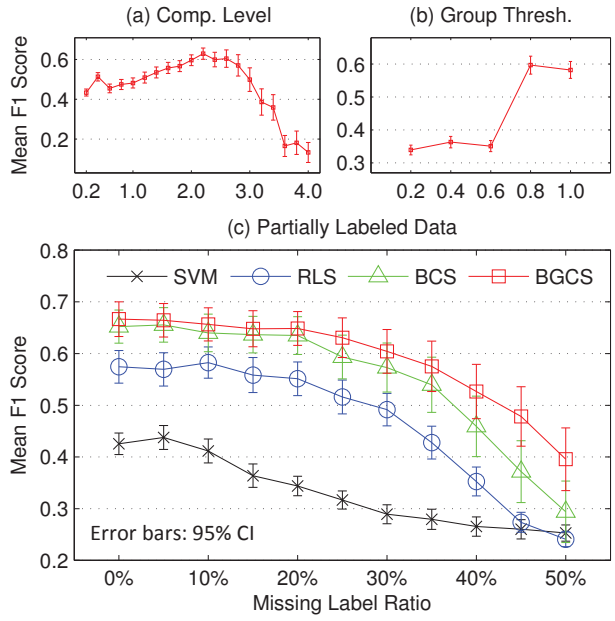

Fig. 4. (a) and (b) show the sensitivity of our model on compression level and group threshold parameters; (c) shows the sensitivity on partially labeled data. See the text for details.

some AUs are much more challenging to detect than others. Figure 4 (a) and (b) show the sensitivity of our model on compression level and group threshold parameters (fixing all other parameters $c = 2.0, \theta = 0.8, \chi = 0.1,$ and $\sigma = 0.01$).

Partially Labeled Data: We evaluate the robustness of the proposed method with respect to partial labels. We simulate missing labels by randomly setting $\eta\%$ of the training data labels to be unobserved; we varied $\eta = [0 : 5 : 50]$. For the SVM and RLS models, the missing labels were randomly assigned to 1 or 0. For the BGCS, only during training, we used element-wise sparsity (testing used group-wise sparsity); we empirically found this performing better, potentially due to less uncertaintly in estimating the sparsity (Gamma) prior distributions.

Figure 4 (c) shows the mean F1-scores for the 24 AUs of different models. These experiments were performed with $c = 2.0, \theta = 0.8, \chi = 0.1, \sigma = 0.01$. The performance increase for the BGCS vs. the BCS is due to the group-wise sparsity during testing (as mentioned, element-wise sparsity is used for training with partially labeled data). This highlights the benefit of considering co-occurrence structure.

Overlapping group structures: One crucial aspect of our approach is the allowance of overlapping group structures, which allows us to deal with action units that occur frequently across multiple groups. Without this capability, every AU will be a part of a single group. As a result, frequently appearing AUs (e.g., AU25) are forced to occur less frequently during inference. The overlapping group structure helps avoid this problem because frequently occurring AUs will be a part of multiple groups and are more likely to be labeled as present (by marginalizing over groups).

As shown in Figure 5, on the CK+ dataset, using BGCS we achieved the highest per-AU F1 scores on each of the top 15 most frequently occurring AUs. The means and standard deviations of F1 scores were: SVM (0.42, 0.21), RLS (0.37, 0.23), BCS (0.45, 0.20), and BGCS (0.55, 0.18). On average, our BGCS performs higher than SVM by .14, RLS by .18, and BCS by .11.

Group-wise AU detection: We measured performance on groups of AUs from the CK+ dataset. A prediction was regarded as correct only if all AUs of a group were detected simultaneously; this reflects the practical application of detecting combinations of AUs (e.g., expressions of emotion or pain). To determine which groups to evaluate, we selected the 12 most frequent AU groups from the CK+ dataset [19].

Figure 6 shows our model significantly outperforming other baselines: the overall F1 scores were 0.31, 0.36, 0.44, 0.48 for SVM, RLS, BCS, and BGCS, respectively. Notably, our model performed particularly well on groups with AU26 (AU25+26 and AU1+2+5+25+26); none of the baselines were able to detect either of the two groups. F1 scores on AU26 alone were quite low for all four models (0.09, 0.0, 0.0, 0.13, respectively) suggesting our simple appearance features (PHOG) may have not been discriminative enough to detect AU26. We believe the group-wise sparsity constraint helped our model outperform other baselines on AU26: the fact that AU26 co-occurred frequently with AU25 encouraged our model to detect them together.

Learning AU groups from [17]: Note that we chose to use an independent dataset [17] to obtain the group structures for two main reasons. First, it helps avoid overfitting: using the same dataset used for training could be problematic because the resulting model may not generalize well. In our preliminary analysis on the CK+ dataset, using the training data for group initialization showed slightly inferior performance (accuracy dropped from 90.7% to 90.4%), which shows the model is overfitted. Second, our approach poses an interesting question of whether it is possible to automatically learn AU groups, based purely on the co-occurrence statistics, that generalize well across different datasets. We show this is possible for the three datasets we have tested, which indeed have different AU co-occurrence structures. The groups in the CK+ dataset are especially different because it contains posed facial expressions.

VI. Conclusions

We have presented a novel method for facial action unit detection that encodes sparsity of facial action units and utilizes the co-occurrence between muscle movements on the face. The benefits of the proposed method include, a principled approach to exploit sparsity and co-occurrence structure in a Bayesian framework, the ability to deal with overlapping groups, superior AU detection performances in both per-AU and per-group settings, and the robustness to missing labels. Experiments show improvements over state-of-the-art for AU detection on posed, acted and spontaneous data. In addition, we presented results across a much larger number of AUs than much of the prior work. In the future, we plan to evaluate our method on the task of AU intensity estimation.

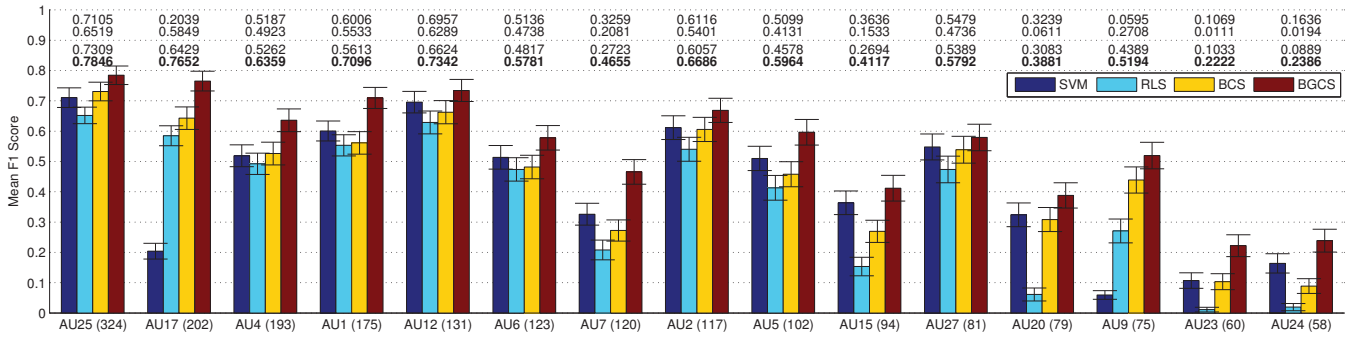

Fig. 5. **AU-wise performance plot:** Mean F1 scores on top 15 frequently occurring AUs of the CK+ dataset [19]. Numbers on the top show mean F1 scores of each model. Numbers in parentheses show the number of times each AU appears in the dataset. The error bar shows 95% confidence interval.

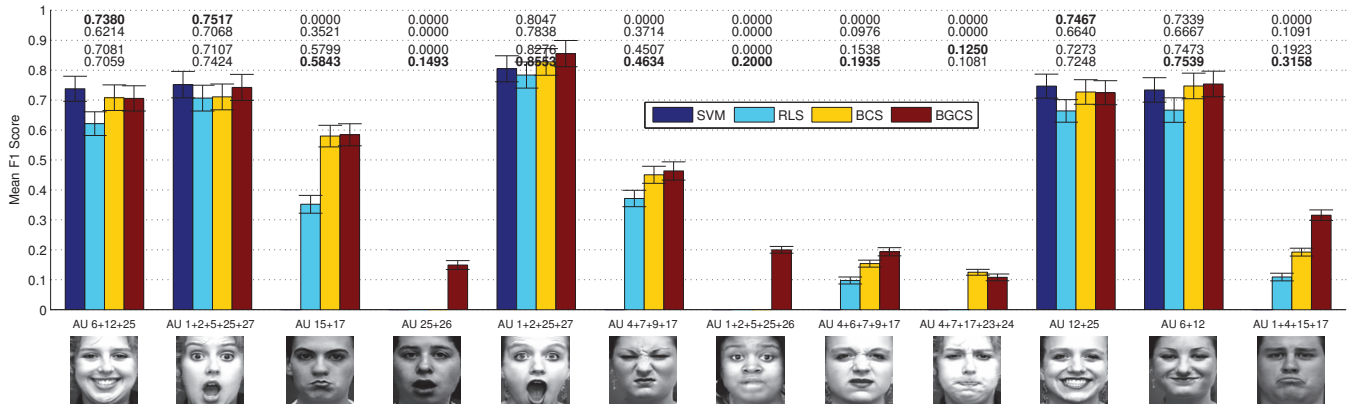

Fig. 6. **Group-wise performance plot:** Mean F1 scores on top 12 frequently occurring AU groups of the CK+ dataset [19]. An instance was counted as correct only if a combination of AUs are detected simultaneously. Numbers on the top show mean F1 scores of each model. Face images © Jeffrey Cohn.

REFERENCES

[1] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Bayesian compressive sensing using laplace priors. *IEEE TIP*, 19(1), 2010.

[2] S. D. Babacan, S. Nakajima, and M. N. Do. Bayesian group-sparse modeling and variational inference. *IEEE TSP*, 2012.

[3] T. Bänziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, 2010.

[4] 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 *CVPR*, 2005.

[5] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.

[6] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8), 2006.

[7] R. Caruana. *Multitask learning*. Springer, 1998.

[8] S. W. Chew, S. Lucey, P. Lucey, S. Sridharan, and J. Conn. Improved facial expression recognition via uni-hyperplane classification. In *CVPR*, 2012.

[9] P. Ekman and W. Friesen. Manual for the FACS. 1977.

[10] P. Ekman and E. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford, 1997.

[11] W. V. Friesen and P. Ekman. EMFACS-7: emotional facial action coding system. *Unpublished, UCSD*, 1983.

[12] P. Garrigues and B. A. Olshausen. Group sparse coding with a laplacian scale mixture prior. In *NIPS*, 2010.

[13] D. Hsu, S. M. Kakade, J. Langford, and T. Zhang. Multi-label prediction via compressed sensing. *NIPS*, 2009.

[14] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38, 2010.

[15] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML*, 2009.

[16] A. Kapoor, R. Viswanathan, and P. Jain. Multilabel classification using bayesian compressed sensing. In *NIPS*, 2012.

[17] K. S. Kassam. *Assessment of emotional experience through facial expression*. PhD thesis, Harvard, 2010.

[18] Y. Li, S. M. Mavadati, M. H. Mahoor, and Q. Ji. A unified probabilistic framework for measuring the intensity of spontaneous facial action units. In *FG*, 2013.

[19] 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 *CVPR*, 2010.

[20] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. In *FG*, 2011.

[21] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE TAC*, 2013.

[22] K. P. Murphy. *Machine learning*. MIT Press, 2012.

[23] S. Raman, T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth. The bayesian group-lasso for analyzing contingency tables. In *ICML*, 2009.

[24] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE PAMI*, 23, 2001.

[25] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE PAMI*, 32(2), 2010.

[26] Y. Tong, W. Liao, and Q. Ji. Inferring facial action units with causal relations. In *CVPR*, 2006.

[27] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *CVPRW*, 2006.

[28] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *FG*, 2011.

[29] M. F. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *Human–Computer Interaction*. Springer, 2007.

[30] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE PAMI*, 31(1), 2009.

[31] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, 2012.