This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing

Copula Ordinal Regression Framework for Joint Estimation of Facial Action Unit Intensity

Robert Walecki[1], Ognjen (Oggi) Rudovic[2], Vladimir Pavlovic[3] and Maja Pantic[1]
[1] Computing Department, Imperial College London, UK
[2] MIT Media Lab, Massachusetts Institute of Technology, USA
[3] Department of Computer Science, Rutgers University, USA

Abstract—

Joint modeling of the intensity of multiple facial action units (AUs) from face images is challenging due to the large number of AUs (30+) and their intensity levels (6). This is in part due to the lack of suitable models that can efficiently handle such a large number of outputs/classes simultaneously, but also due to the lack of suitable data the models on. For this reason, majority of the methods resort to independent classifiers for the AU intensity. This is suboptimal for at least two reasons: the facial appearance of some AUs changes depending on the intensity of other AUs, and some AUs co-occur more often than others. To this end, we propose the Copula regression approach for modeling multivariate ordinal variables. Our model accounts for $ordinal$ structure in output variables and their $non\text{-}linear$ dependencies via copula functions modeled as cliques of a conditional random fields. The copula ordinal regression model achieves the joint learning and inference of intensities of multiple AUs, while being computationally tractable. We demonstrate the effectiveness of our approach on three challenging datasets of naturalistic facial expressions and we show that the estimation of target AU intensities improves especially in the case of (a) noisy image features, (b) head-pose variations and (c) imbalanced training data. Lastly, we show that the proposed approach consistently outperforms (i) independent modeling of AU intensities and (ii) the state-of-the-art approach for the target task and (iii) deep convolutional neural networks.

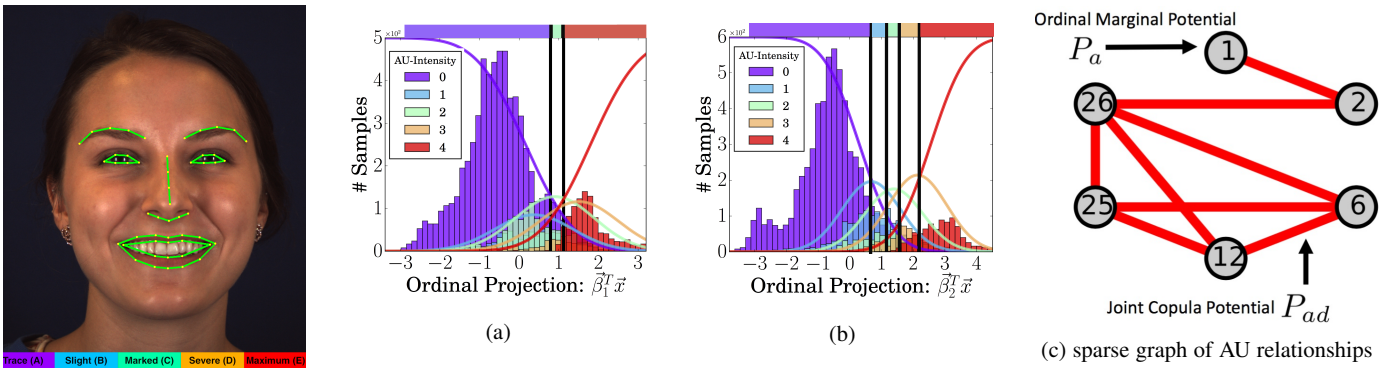
1 Introduction

Human facial expressions are typically described in terms of variation in configuration and intensity of facial muscle actions defined using the Facial Action Coding System (FACS) [8]. Specifically, the FACS defines a unique set of 30+ atomic non-overlapping facial muscle actions named Action Units (AUs) [37]. It also provides rules for scoring the intensity of each AU in the range from absent to maximal intensity on a six-point ordinal scale, denoted as *neutral*<A<B<C<D<E. Thus, using FACS, human coders can manually code nearly any anatomically possible facial expression, decomposing it into specific AUs and their intensities. However, this process is tedious and error-prone due to the large number of AUs and the difficulty in discerning their intensities [38]. On the other hand, automated estimation of the AU intensity is challenging for many reasons such as the subject-specific facial morphology and expressiveness level [45], as well as the changes in lighting and the head-pose variation.

Co-occurrences of the intensity levels of different AUs are another important factor that affects their coding/automated estimation. For instance, the criteria for intensity scoring of AU7 (lid tightener) are changed significantly if AU7 appears with a maximal intensity of AU43 (eye closure), since this combination changes the appearance as well as timing of these AUs [8]. Furthermore, co-occurring AUs can be non-additive, in the case of which one AU masks another, or a new distinct set of appearances is created [8]. As an example of the non-additive effect, AU4 (brow lowerer) appears differently depending on whether it occurs alone or in combination with AU1 (inner brow raise). When AU4 occurs alone, the brows are drawn together and lowered, while in AU1+4, the brows are drawn together but are raised due to the activation of AU1. This, in turn, significantly affects their appearance. Moreover, some AUs are often activated together, e.g. AU12 and AU6 in the case of smiles, but with different intensities depending on the type of smile (e.g., genuine vs. posed [2]). Therefore, modeling dependencies among (the intensities of) multiple AUs is expected to result in models that are more robust to noisy features and imbalanced training data, leading to a more accurate estimation of the target AU intensities [30], [47].

To date, most of the work on automated analysis of AUs has focused on detection of the presence/absence of AUs (e.g., [7], [34], [38]) instead of their full range intensity estimation. Furthermore, few methods attempted joint modeling of AUs activations (e.g., [9], [31], [65], [79]). However, these methods can deal only with the $binary$ classification problems, and, thus, are not applicable to the joint estimation of intensity of multiple AUs. Because the AU intensity estimation is a relatively new problem in the field, few works have addressed it so far. Most of these works perform independent estimation of the AU intensity using either classification-based approach [37], [39], [47] or regression-based approach [21]. To the best of our knowledge, the only methods that attempt joint estimation of multiple AUs intensity are reported in [22], [30], [50]. The methods [30], [50] perform a two stage joint modeling of AU intensity. Specifically, in [50], the scores of the pre-learned regressors, such as Support Vector Regression, are fed into a set of Markov Random Field trees, used to model dependencies of subsets of AUs. Similarly, [30] models AU dependencies using a Dynamic Bayesian Network (DBN) approach, which feeds as inputs the AU-specific spectral regressors. The current state-of-the-art approach for the joint modeling of the AU intensity [22] formulates a generative MRF model, called Latent Tree (LT). In contrast to the two works mentioned above, this method can deal with the highly noisy and missing input features due to its generative component. Nevertheless, there are several critical limitations of the proposed approaches. The model outputs in [50] are treated as continuous, despite the fact that the intensity levels are defined on an ordinal (discrete) scale. Furthermore, in performing the two-stage learning, [30], [50] fail to allow the input features to influence the learned AU dependencies. Although defined in a probabilistic manner, the LT approach [22] relies on a set of heuristics for the model to be computationally tractable for more than few AUs. Also, Explicitly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing



(a) (b) (c) sparse graph of AU relationships

Fig. 1: Joint modelling of AU intensities. The marginals (a, b) are computed AU separately and are directly used in independent models for predictions. The intensities thresholds are shown in solid black lines. Note that the AU1 marginal model here is incapable of predicting levels 1&3. Yet, due to the strong association between AU1&2 , the joint model overcomes this by learning a joint distributions, described by copulas, each pair of AUs. This joint distribution is represented by a sparse graph of AU relations in which the edge potentials, defined by bivariate copula functions, and the node potentials, defined by ordinal regressors, are jointly learned.

modeling the relations between co-occurrences has been addressed recently for binary detection (e.g., for object detection [32]), and not for multi-level intensities. The proposed Copula Ordinal Regresion (COR) model addresses these limitations. We model the AU intensity relations by allowing them to be *non-linearly* related - in contrast to present models that account only for linear dependencies. We do so by means of copula functions, known for their ability to capture highly non-linear dependencies through a simple parametrization. The notion of the copula functions has previously been explored for modeling of structured output [71] for AU intensity estimation. However, this model has two limitations. First, the AU intensity is modelled in a homoscedastic manner. More precisely, the full range of AU intensities is modelled with the same correlation parameter. This is suboptimal since AU dependencies can be very different for facial expressions with high and low intensity. Secondly, the potentials of the CRF are optimized independently and the inference is performed globally. This makes the optimization efficient but higher-order dependencies are ignored by that model.

Contributions. To address the primary challenge of computational modeling of variable and complex dependencies that exist among intensities of multiple AUs, we propose the Copula Ordinal Regression model for joint AU intensity estimation. Specifically, we propose to use the powerful framework of copula functions [58] to efficiently model dependencies of intensities among AUs. Copula functions generalize the notion of linear correlation to more flexible dependency structures specified using simple parametric functional families (copula families). The key advantage of copula models is that they retain representational and computational efficiency by decoupling the modeling of dependencies from the modeling of marginal densities, as detailed in Sec.3.2. The basic idea is that one starts with state-of-the-art independent (marginal probability) AU models and then captures the intrinsic AU dependence (joint probability) through copula functions, while guaranteeing that the marginals remain unaltered. This presents a distinct advantage over all previously surveyed models that tightly couple the marginal and joint model specification/estimation, resulting in often intractably complex models.

Even though copulas model dependencies using compact parametric functions, it is still necessary to estimate their parameters from data. To this end, we propose a new Conditional Random Field (CRF) model in Sec.3.2 and the accompanying learning and inference strategies in Sec.3.5. The CRF-based model considers sparse, graph-induced, cliques of AUs (inferred from data and illustrated in Fig.5), where dependencies in each clique are modeled using an independent copula model. The joint CRF model is then estimated using a new, efficient block descent algorithm that intuitively combines optimization of dependencies (copula association parameters) with learning of independent marginal model parameters (the intensity levels of each AU from the corresponding covariates, i.e., the locations of a set of fiducial facial points). Furthermore, we introduce the heteroscadastic association for Copula regression. The association is directly learned from the data and depends on the level of intensity of the target AUs. To avoid the typically challenging evaluation of the CRF partition function, we propose to use piece-wise optimization [32], [62], of the CRF. In this way, the higher order dependencies are considered while the optimization kept tractable. The joint inference in this model is then accomplished using a fast loopy belief approximation method on the learned CRF model. Below we emphasize the main contributions of the proposed work:

- We propose a novel structured CRF model for joint learning of multiple ordinal outputs. The data structure is seamlessly embedded via an undirected graphical model, capturing the ordinal structure in AU intensity levels via ordinal unary cliques, and non-linear dependencies between the network outputs via the copula binary cliques. We show that this model better estimates the intensities of the target AUs, especially from scarce and highly im-balanced data.
- The proposed work introduces novel methodology for multi-class multi-output ordinal learning. While this approach relies on ideas from ordinal modeling, it uses the copula framework to tackle the challenging problem of intensity estimation of multiple AUs. Furthermore, we propose the pice-wise CRF optimization to achieve efficient learning and inference. We also demonstrate the robustness against noisy image features and head-pose variations by evaluating the model on noisy datasets with a relative low resolution (Shoulder Pain) and by testing and comparing the models on only nonfrontal images in the test data.
- We show that the proposed approach significantly outperforms the state-of-the-art approaches ([22], [23], [50], [71]) and deep CNNs ([13], [43], [80]) to the target problem in the experiments on three challenging benchmark datasets for modeling of AU intensity levels, FERA2015 [68], DISFA [39] and Shoulder Pain [35].

2 Related Work

Over the past decades, there has been extensive research in computer vision on facial expression analysis [52], [53], [70]. Here we will

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing

present a brief review of the previous work on AU intensity estimation and other related methods that are applicable to this problem. As already mentioned before, most past work addresses the problem of either facial affect detection or AU detection. The problem of automatic AU intensity estimation has been tackled by few works only recently. Predecessors of this line of research were works on temporal dynamics of AUs. These works aimed to encode tempral segments - onset, apex, and offset of AUs - rather than actual intensity of AUs. examples of such works are [19], [20], [51], [69].

AU intensities estimation can be divided into traditional shape and appearance based models and models that rely on deep convolutional neural networks. Recently, it was shown that in order to achieve state-of-the-art results in a series of important computer vision applications, such as face recognition/verification and facial expression analysis, it is important to provide an enhanced representation of the face that contains the locations of several key facial landmarks [53]. However, recent advances in deep neural networks (DNN), and, in particular, convolutional models (CNNs) [13], have allowed to completely remove or highly reduce the dependence on physics-based models and/or other pre-processing techniques, by enabling the 'end-to-end' learning in the pipeline directly from input images. While the effectiveness of these models has been demonstrated on many general vision problems [27], [61], [63]. These models have been applied on baseline tasks such as expression recognition AU detection [25], [33], [80] and AU intensity estimation [13] have been investigated. All of them, however, follow the traditional 'blind deep learning' paradigm that relies on large labeled training datasets (e.g., 100K+ samples in [44]). Yet, in the facial data domain, obtaining accurate and comprehensive labels is typically prohibitive. For instance, it takes more than an hour for an expert annotator to code AUs intensity for 1 minute of face video. Even then, the annotations are highly biased and have low inter-annotator agreement. Coupled with large variability in imaging conditions, facial morphology, dynamics of expressions, this has resulted in the lack of suitable large datasets for effective deep model learning. Therefore, not surprisingly, appearance based models are still superior to CNNs for the task on AU intensity estimation. In what follows, we will review the related work for both, regression-based and classification-based methods for AU intensity estimation. We will also include the work on CNNs and compare the performance of those models.

Classification

The majority of the existing works attempt to recognize AUs independently. This is done by either static of dynamic models. Static models typically tackle the problem as discriminative classification, in which each video frame is evaluated independently. Examples are: Neural Network [24], Adaboost [4], and SVMs [36]. Dynamic models also capture the temporal transition between adjacent frames in a sequence. For instance, Dynamic Bayesian Network (DBN) with appearance features [66]. Other variants of DBN are based on Hidden Conditional Random Fields for AU detection [72]. While the methods above are used for AU-detection, it is however needed to treat the AU intensity estimation problem as 6-class classification. For example, the authors of [38] employed six one-vs.-all binary SVM classifiers. Alternatively, a single multi-class classifier (e.g. ANN, Boosting or SVM) could be used. The second Facial Expression Recognition and Analysis Challenge (FERA15 [68]) proposed a sub-challenge for AU intensity prediction, where most of the proposed systems are based on independent multi-class classifier. For example, [41] applied feature fusion (appearance and geometry) with a multi-kernel Support Vector Machine. Similarly, [17] applies decision-level fusion strategies for AU intensity estimation. Instead, [3] focused on robust features extraction with the goal to improve the performance for cross domain experiments.

Regression

AU intensity estimation is nowadays often posed as a regression problem. Regression methods penalise incorrect labelling proportionally to the difference between ground truth and prediction. Such consideration of the labels is absent in most of the classification based methods. Examples include Support Vector Regression [14], [18], [54]. [54] aims to learn an regression based model by applying logistic regression on SVM scores. Instead, [21] used Relevance Vector Regression to obtain a probabilistic prediction. [18] use the confidence of a (binary) classifier to estimate AU intensity. The rationale is that the lower the intensity is, the harder the classification will be. For example, Doubly Sparse Relevance Vector Machine (DSRVM) [23] was proposed for pain intensity estimation. In this work, features are locally extracted from a pre-defined grid of rectangular regions in face images registered in frontal pose. Hence, this technique is not suitable for images with large head pose variations, typical of real world scenarios, since the 2D registration process unavoidably induces pixel artifacts and texture discontinuities. Furthermore, some researchers are critical of the grid-based feature extraction, suggesting that the sub-regions are not necessarily well aligned with meaningful facial features [16].

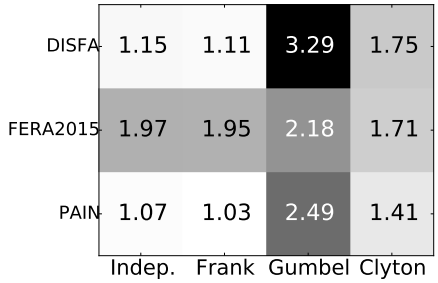
Multi-Output Detection

A common limitation of most above mentioned independent methods is that they construct single classifier or regressors that ignore the relations among AUs. Multi-Output-Detection (MLD) attempts to learn robust classifiers by exploiting efficiently the output dependencies. For an extensive overview, the reader is referred to [60], [67]. MLD can also be grouped in two categories: binary and multi-class classification. Only a few work in MLD addresses the problem of joint AU-detection. For example, [76] investigate the multi-label AU detection problem by embedding the data on low dimensional manifolds which preserve AU dependencies. In [65], the authors propose to extract the most discriminative features for each AU individually and then jointly model the AUs in a multi-task framework. In similar fashion, [82] uses a Bayesian network to model the co-existent and mutual-exclusive semantic relations among AUs from the labels. The work of [73] applied a Restricted Bolzman Machine for classification in which the top layer captures the global relationships among the AUs. [48] aims to train a Multilayer perception classifier with landmarks as targets and expressions as the source data. In this work, AUs can be seen as jointly modeled latent states. These methods have the common problem that joint events occur less frequently which results in an sparse label space. To address this problem, [59] proposed compressed sensing and group-wise sparsity priors on AUs. Alternatively, joint patch learning for AU detection has been recently used to jointly learn dependencies among groups of AUs [64], [79], [81] While applicable for joint AU-detection, the methods above do not perform multi-class classification.

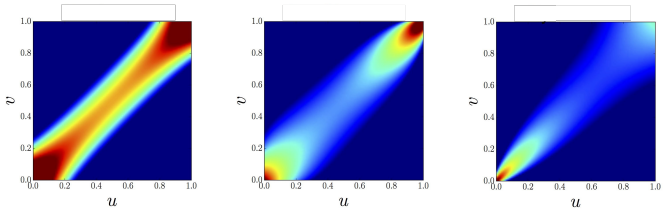
Multi-Output Classification

In contrast to the methods above, the proposed COR Framework can be seen as a **Multi-output-multi-class** (MOMC) learning approach, where the relations of different pairs of AUs are learned directly with the copula model. Only very little related work exists for MOMC AU intensity estimation. Recently, [42] proposed multi-task learning for AU intensity estimation where a metric with shared properties among multiple outputs is learned, resulting in less parameters to learn and a reduction of over fitting. However, this method is not directly applicable to larger datasets since the complexity scales squared with the number of samples. In similar fashion, [77] derived the back-propagation algorithm of the neural networks for multi-label classification. Both methods have the problem that as the dimensionality and the number of outputs increase, they become computationally hard

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing



(a) goodness of fit from different copulas



(b) Frank (c) Gumbel (d) Clyton

Fig. 2: Average negative conditional log likelihood (NCLL) for copula model on different datasets. 2 for all pairs of AUs using different copulas on all three datasets.

to solve and common structures are more difficult to identify. For example, [13] proposed a deep convolutional neural network (CNN) for joint AU-intensity estimation. CNNs have been recently very successful in very different visual tasks, especially for detection. However, the performance obtained by this DNN is worse than the baseline model using a simple SVM classifier on geometrical features. One reason for this low performance is the absence of fully annotated data which makes it difficult to apply DNNs/CNNs successfully. In order to deal with little data and spares co-occurrence, [30] employs a two-step approach: the outputs of SVMs learned for each AU independently are used to model AU dependencies via Dynamic Bayesian Network (DBN). Despite of being simple to train, two stage approaches result in suboptimal solutions [50]. In similar fashion, [50] constructs a MRF-tree-like model for joint intensity estimation of AUs. In this work, the dependency graph is limited to a tree structure which does not represent true dependencies among all features. Moreover, this approach also uses two steps learning - by first obtaining the intensity scores for each AU independently, followed by graph optimization. This is again suboptimal as it results in loss of information from the feature level. More recently, [22] proposed Latent-Trees (LT) for joint AU-intensity estimation. LT is a probabilistic model in which a tree-structure is learned by maximizing the log-likelihood of training data while maintaining model complexity low. In comparison to single-target-regression methods, LT have better generalization capabilities and it is more effective due to the learned structure that captures higher-order dependencies among the high-dimensional input features and multiple target AU intensities.

To the best of out knowledge, the work proposed here is the first to estimate AU intensities jointly using a single objective function. By applying ordinal constrains on the node potential and describing the edges using bivariate copula functions we prevent the model to fall in a local minimum and reduce the number of parameter. This results in the state-of-the-art model for AU intensity estimation.

3 Methodology

Let us denote the training set as $\mathcal{D} = \{\mathbf{Y}, \mathbf{X}\}$. $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N]^T$ is comprised of N instances of multivariate outputs stored in $\mathbf{y}_i = \{\mathbf{y}_i^1, \ldots \mathbf{y}_i^q, \ldots \mathbf{y}_i^Q\}$, where Q is the number of AUs, and $\mathbf{y}_i^q$ takes one of $\{1, ..., L^q\}$ discrete intensity levels of the q-th AU. Furthermore, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N]^T$ are input features (e.g., facial points) that correspond to the combinations of labels in $\mathbf{Y}$. Thus, our goal is to simultaneously estimate the combination of the intensity levels $\mathbf{y}^q$ of Q AUs, given the facial features $\mathbf{x}$. In what follows, we first introduce the ordinal regression framework for modeling single output ($Q = 1$). We then introduce the copula framework for modeling joint distributions, and formulate our model for joint learning and inference of intensity levels of multiple AUs.

3.1 Ordinal Regression

Let $l \in \{1, \ldots, L^q\}$ be the ordinal label for the intensity level of the q-th AU. In the ordinal regression framework notation [1], we define the latent projection $y_*^q \in \Re$ as a function of covariates x, and then relate this latent projection to the ordinal level (y^q) through the threshold bounds:

$$y_*^q = \beta^{\mathbf{q}}\mathbf{x}^{\mathbf{T}} + \varepsilon^{\mathbf{q}}, \mathbf{y}^{\mathbf{q}} = \mathbf{l} \text{ iff } \psi_{\mathbf{l-1}}^{\mathbf{q}} < \mathbf{y}_*^{\mathbf{q}} \leq \psi_{\mathbf{l}}^{\mathbf{q}}, \quad (1)$$

where $x \in \Re^D$, $\beta^{\mathbf{q}}$ is the ordinal projection vector, ψ_l^q is the lower bound threshold for count level l ($\psi_0^q = -\infty < \psi_1^q < \psi_2^q ... < \psi_{L-1}^q < \psi_L^q = +\infty$). The error (noise) terms ε^q capture the idiosyncratic effects of all omitted variables for the q-th AU. They are assumed to be identically distributed across the intensity levels, each with a univariate continuous marginal distribution function $F(z^q) = \Pr(\varepsilon^q < z^q)$. In the case of the normal distribution with zero mean and variance $(\sigma^q)^2$, the marginal distribution function is defined as the normal cumulative density function (cdf) $F(z^q) = \Phi(z^q) = \int_{-\infty}^{z^q} \mathcal{N}(\xi; 0, 1)d\xi$. Then, classification in ordinal regression models is performed using the following $ordinal$ likelihood [1]:

$$l^* = \underset{l=1\ldots L}{\operatorname{argmax}} \Pr(y^q = l|\mathbf{x}) = \underset{\mathbf{l=1\ldots L}}{\operatorname{argmax}} \mathbf{F}(\mathbf{z_l^q}) - \mathbf{F}(\mathbf{z_{l-1}^q}), \quad (2)$$

where $z_k^q = \frac{(\psi_k^q - \beta^{\mathbf{q}}\mathbf{x}^{\mathbf{T}})}{\sigma^q}$ are the cumulative probits. The model parameters are then stored in $\varphi^q = \{\psi_1^q, \psi_2^q, \ldots, \psi_{L-1}^q, \beta^{\mathbf{q}}, \sigma^{\mathbf{q}}\}$.

3.2 Copula Model

A copula is a method for generating a stochastic dependence relationship in the form of a multivariate distribution of random variables with pre-specified marginals [55]. Formally, a copula $C(u^1, u^2, \ldots u^Q)$: $[0, 1]^Q \rightarrow [0, 1]$ is a multivariate distribution function on the unit cube with uniform marginals [74]. Several copulas have been proposed and they can be grouped into two type of copula families that are Gaussian and Achimedean copulas. The Gaussian copula is constructed from a multivariate normal distribution by using the probability integral transform of the multivariate gaussian. While there is no simple analytical formula for the copula function, it can be upper or lower bounded, and approximated using numerical integration. Archimedean are widely used in applications due to their simple form and closed form solution. There are three Archimedian copulas [12] in common use: Clayton, Frank and Gumbel. The Clayton copula (Fig.2d) is an asymmetric Archimedean copula, exhibiting greater dependence in the negative tail than in the positive. For two dimensions, this copula is given by:

$$C_\theta^{\text{Clyton}}(u, v) = max((u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}) \quad (3)$$

The Frank copula (Fig.2b) is a symmetric Archimedean copula and has a diagonal probability density. It can be expressed as:

$$C_\theta^{\text{Frank}}(u, v) = -\frac{1}{\theta}\ln\left(1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{\exp(-\theta) - 1}\right) \quad (4)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing

The Gumbel (Fig.2c) copula is an asymmetric Archimedean copula, exhibiting greater dependence in the positive tail than in the negative. This copula is given by:

$$C_\theta^{\text{Gumbel}}(u,v) = \exp\{-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{\frac{1}{\theta}}\} \tag{5}$$

Copulas are selected by choosing a particular member of a given family of Archimedean copulas to fit a data set. Archimedean Copulas are described in a closed form analytic expression for the joint probability of choice across observational units, using a standard and direct maximum likelihood inference procedure. The Gaussian has no closed form solution and the joint probability is normally approximated (by MCMC). For this reason, work we focus for this Framework on Archimedian Copulas.

The main idea of copulas related to that of histogram equalization: for a random variable y^q with (continuous) cdf F, the random variable $u^q := F(y^q)$ is uniformly distributed on the interval $[0,1]$. Using this property, the marginals can be separated from the dependency structure in a multivariate distribution [5]. This is given by Sklar's theorem [58].

Theorem 1 (Sklar, 1973) *Given u^q random variables with cdfs F_i, $q = i,\dots,Q$, and joint distribution F_i on $y^i,\dots,y^Q$, there exist a unique copula C such that for all u^q:*

$$C\left(u^1,\dots,u^Q\right) = F\left(F_1^{-1}(u^1),\dots,F_Q^{-1}(u^Q)\right) \tag{6}$$

Conversely, given any distribution functions $F_1,...,F_Q$ and copula C,

$$F(y^1,\dots,y^Q) = \mathrm{C}(F_1(y^1),\dots,F_Q(y^Q)), \tag{7}$$

is a Q-variate distribution function on $y^1,\dots,y^Q$ with marginal distribution functions F.

This result allows us to construct a joint distribution by specifying the marginal distributions and the dependency structure *separately* [5]. This offers one the critical flexibility necessary for any multivariate output context: it is possible to simultaneously model complex marginal densities with potentially arbitrary multivariate output dependency structures without the need to specify the two in some complexly intertwined, hard-to-interpret and hard-to-learn model. Note that while the copula representation separates the two contexts (marginal and joint) the two remain tied through Eq. 6.

When the random variables are discrete, as is the case with the AU intensity levels, only a weaker version of **Theorem 1** holds: there always exists a copula that satisfies Eq. 7, but it is no longer guaranteed to be unique [58]. Nevertheless, we can still construct the joint distribution for discrete variables as:

$$\begin{aligned}
&\Pr(y^1 = l^1,\dots,y^Q = l^Q) = \\
&\Pr(\psi_{l^1-1} < y_*^1 < \psi_{l^1},\dots,\psi_{l^Q-1} < y_*^Q < \psi_{l^Q}) \\
&= \sum_{c_1=0}^{1}\cdots\sum_{c_Q=0}^{1}(-1)^{c_1+\dots+c_Q}F(z^1_{l^1-c_1},...,z^Q_{l^Q-c_Q}) \\
&= \sum_{c_1=0}^{1}\cdots\sum_{c_Q=0}^{1}(-1)^{c_1+\dots+c_Q}C_\theta(u^1_{l^1-c_1},...,u^Q_{l^Q-c_Q})
\end{aligned} \tag{8}$$

where $u^q_{l^q-c_q} = F(z^q_{l^q-c_q}), c_q \in \{0,1\}$, is defined in Sec.3.1, and θ are the copula parameters, as defined below. It is important to note two critical aspects here. First, Eq. 8 captures dependency structures among the discrete outputs by correlating their error terms $\varepsilon^1,\dots,\varepsilon^Q$ via the copula. Second, the joint density model induced by the copula is conditioned on the covariates x, as explained in [12], i.e., $F(y^1,\dots,y^Q) \leftarrow F(y^1,\dots,y^Q|x)$. This, in contrast to the models in [30], [50] that rely solely on the AU labels, allows the covariates to directly influence the dependence structure of AUs.

Under this formulation, the probability of a particular label combination $\mathbf{y}$ is determined by the volume of the axis-parallel hyperrectangular subregion of $[0,1]^Q$ induced by vertices $(u^1_{l^1},\dots,u^Q_{l^Q})$ and $(u^1_{l^1-1},\dots,u^Q_{l^Q-1})$ corresponding to that label combination. For the copula introduced in Eq. 8, this involves evaluation of 2^Q cdfs. As an example, for $Q = 2$ the model this reduces to:

$$\begin{aligned}
&\Pr(y^1 = l^1, y^2 = l^2) = F(z^1_{l_1}, z^2_{l_2}) \\
&+F(z^1_{l_1-1}, z^2_{l_2-1}) - F(z^1_{l_1-1}, z^2_{l_2}) - F(z^1_{l_1}, z^2_{l_2-1})
\end{aligned} \tag{9}$$

This evaluation becomes computationally expensive and impractical for $Q > 5$ due to the number of cdfs (2^Q) that need be evaluated. In Sec. 3.3, we propose a computationally more astute model, which avoids the exponential explosion induced by arbitrary Q.

One specific benefit of copulas is that they can model different forms of (non-linear) dependency using simple parametric models for $C(\cdot)$. In this paper, we limit our consideration to the commonly used Frank copula [11] from the class of Archimedean copulas. The dependence parameter $\theta \in (-\infty,+\infty)\backslash\{0\}$, and the perfect positive/negative dependence is obtained if $\theta \to \pm\infty$. When $\theta \to 0$, we recover the set of independent ordinal models of Eq.2 correlating to $C(u,v) = u \cdot v$. Although various copula functions (e.g., Clayton, Gumbel, etc.) are available for modeling different dependence structures, we choose Frank copula in this paper for two reasons. First, it has a simple closed-form, in contrast to, e.g. , the Gaussian copula [5], which, in general, requires the intractable computation of multivariate Gaussian cdfs. Second, Frank copula is particularly suitable for the target task as it allows modeling of both positive and negative dependencies, while also capturing dependency in both the left and right tails (i.e., when different AUs are activated either at low intensity, or at high intensity levels together). However, we also provide qualitative results for the Clython, Gumbel and Independent Copula.

3.3 Copula Ordinal Regression

As mentioned in Sec.3.2, the joint modeling of multiple AUs using the model in Eq.8 is possible. However, this becomes prohibitively expensive as the number of outputs (i.e., AUs) increases. For instance, for 10 AUs and 6 intensity levels, as commonly coded in face datasets, this would involve 6^{10} evaluations of the copula function. We mitigate this by approximating the learning of the joint pdf in Eq.8 using the bivariate joint distributions capturing dependencies of AU pairs.

To this end, we use the Conditional Random Field (CRF) [29] framework. Formally, we introduce a random field with an associated graph $\mathcal{G} = (V,\mathcal{C})$, where nodes $v \in V, |V| = Q$, correspond to individual AUs and cliques $c \in \mathcal{C}$ correspond to subsets of dependent AUs modeled using the copula functions. The joint probability distribution of Q intensity random variables is then defined as:

$$P(\mathbf{y}|x,\Omega) = \frac{1}{Z}\prod_{c\in\mathcal{C}}\Psi(\mathbf{y}_c|x) \tag{10}$$

$$\Psi(\mathbf{y}_c|x) = \exp\left[\sum_{r\in V}\ln\phi_r(y_i^r|x_i) + \sum_{(rs)\in E}\ln\phi_{rs}(y_i^r,y_i^s|x_i)\right] \tag{11}$$

where Z is the partition function, $\mathbf{y}_c$ is the subset of random variables in clique c, $\Psi(\cdot)$ is the conditional potential on the labels in this clique, explained below, and $\Omega = \{\vartheta,\theta\}$ are the model parameters.[1] In this setting specifically, we only consider unary and binary cliques, modeling individual independent AUs and pairs of AUs. In other words, $\mathcal{C} = V \cup E$, where E is the set of edges in $\mathcal{G}$. Hence,

$$\Psi(\mathbf{y}_c|x) = \begin{cases} \phi_r \to \Pr(y^r|x), & \begin{matrix} c = r \in V \\ \text{unary clique} \end{matrix} \\ \phi_{rs} \to \Pr(y^r,y^s|x)^\gamma, & \begin{matrix} c = (r,s) \in E \\ \text{pairwise clique} \end{matrix} \end{cases} \tag{12}$$

[1]For simplicity, we often drop the dependency on Ω in notations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing

where the unary term is the traditional independent AU ordinal regression model defined in Sec. 3.1 and the pairwise term is specified in Eq. 9. Note that the unary terms depend only on the ϑ_r parameters of the ordinal regression model, while the edge potentials depend also on the copula association parameter θ_{rs} that models the dependency of (r, s) pair of outputs. Furthermore, the weight γ is chosen so as to balance the magnitude of the cliques.

While modeling only bivariate distributions may seem a natural way of representing the joint distribution, we model also the marginals via the unary potentials for two reasons. First, while the marginals focus on independent classification of target AU intensity, the bivariate copulas focus on encoding the dependence between the intensity levels of two AUs. Thus, by including the copulas in the potential function, a more discriminative classifier for the AU intensity levels is expected. Second, in the case when there is no dependence between AUs, in an ideal case $\theta_{rs} \rightarrow 0$, and Frank copula converges to the independence copula [11]. Yet, due to numerical instability, parameter estimation can be fragile in this case, leading to poor performance of the learned classifier. We control this by having the marginals in the unary potentials.

The most critical aspect in evaluation of the joint distribution in Eq. 11 is computation of the partition function. This is an np-complete problem, and thus, exact inference in general case is intractable. This is true in our case as it involves the integration over all possible AUs and their intensity levels, i.e, typically 6^{10} computations. We decompose the graph into smaller subgraphs and apply exact inference in an iterative manner on each. By using the notion of the negative log likelihood, our learning objective can be written as:

$$NCL = -\sum_{i=1}^{N}\{\log(\Psi(\mathbf{y}_i|x)) - \log(Z)\} \tag{13}$$

Here, N is the number of training instances.

3.4 Estimation of the AU pairs.

We propose the Copula Framework in 3 different configurations:

3.4.1 COR-F

Copula Ordinal Regression with a fully connected MRF is the most general model and serves as a baseline for the more advanced configurations. The graph is build by $Q \times (Q-1)/2$ bivariate copulas and Q nodes.

3.4.2 COR-IT

Modeling all bivariate copulas is impractical as not all AU exhibit a dependence pattern (e.g., AU16 (lower lip depressor) and AU17 (chin raiser) do not co-occur). In CRF and MRF models, the cliques (i.e., the edges) are typically determined from the precision matrix rather than from the correlation matrix S. This is because the precision matrix unravels partial correlations among the variables, while the correlation matrix focuses on marginal correlations [15]. Important advantage of using partial correlations to infer AU dependencies is that, in contrast to marginal correlations, AUs that are correlated through another AU are ignored, therefore, avoiding the redundant modeling. We apply the graphical lasso [10] on the precision matrix from the labels to get a sparse graph containing only bivariate potentials of AU pairs with a certain dependence. The lasso is defined as follows:

$$(\Upsilon, \tilde{S}) = \min_{\Upsilon \succ 0} -\ln\det(\Upsilon) + tr(S\Upsilon) + \kappa\|\Upsilon\|_1, \tag{14}$$

where κ is the regularization parameter.[2] Finally, the edge set E is defined by keeping the edges satisfying the condition: $E = \{(r, s) : |\Upsilon_{r,s}| > \delta\}$. $\delta = 0.05$ is chosen so that only the pairs of AUs with strong partial correlations are kept, resulting in a model with significantly fewer parameters [40] depicted in Fig. 5. The aim is to learn a graph directly from the features rather than predefining it from the labels. This reduces the number of model parameters during training by not accounting for the 'weak' dependencies among AUs.

[2]We used the glasso Matlab code from [10].

3.4.3 COR-HIT

The COR model mentioned above are limited to a homoscedastic association over the full intensity range. However, AU dependencies varies significantly when showing the same facial expression with different intensity. The heteroscadastic model (COR-HIT) tackles this problem by extending the linear ordinal projection with heteroscedasticity. Another limitation of the COR-IT is that the graph structure is fixed after initialization. In the COR-HIT model, we use a sparse l1 regularization on the edge potentials and learn the graph structure iteratively and directly form the training data. The association for each pair of AUs is computed as follows:

$$\theta(\vec{\mathbf{x}}) = x^T h + h_0 \tag{15}$$

The COR-HIT model is then derived by updating the association parameter θ with the feature dependent function $\theta(\vec{x})$. Lastly, the parameter $\vec{h}$ and h^0 are jointly learned by minimizing the objective function.

3.5 Learning and Inference

Algorithm 1: Copula Ordinal Regression Learning

Input: Training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$
Result: Model parameters $\Omega = \{\varphi, \theta\}$
initialization;
$\forall (r, s) \in E \rightarrow \theta_{rs} = sign(corr(y^r, y^s))$
$\forall r \in V \rightarrow \varphi_r = \arg\min_{\varphi'} - \sum_{i=1}^{N \in AU_r} \mathcal{L}_r(\varphi)$
while *score increases* **do**
  φ-step: optimize φ_r using subset of active frames
  for $r \in V$ **do**
    $\mathcal{L}_r(\varphi_r) = \sum_i \mathbb{1}_{[y_i^r > 0]} \ln \phi_r(y_i^r | x_i, \varphi_r)$
    $\varphi_r = \arg\min_{\varphi'} - \mathcal{L}_r(\varphi) + \lambda_r R_{\varphi_r}$
  θ-step: optimize θ_{rs} using jointly active frames
  for $(r, s) \in E$ **do**
    $\mathcal{L}_{rs}(\theta_{rs}) = \mathbb{1}_{[y_i^r, y_i^s > 0]} \ln \phi_{rs}(y_i^r, y_i^s | x_i, \theta')$
    $\theta_{rs} = \arg\min_{\theta'} - \mathcal{L}_{rs}(\theta') + \lambda_{rs} R_{\theta_{rs}}$
    Pruning-step: remove edges with low weights
    if $w_{rs} \sim 0$ **then**
      $E \leftarrow E \setminus \{e_{rs}\}$
  Evaluation-step:
  Use Alg. 2 to get predictions and compute ICC on validation set. $\mathbb{1}$ is the indicator function it is 1 if the argument to that function is true.

The parameter optimization in the model is performed by minimizing NCL (Eq.13) w.r.t. Ω. For this, we employ the Conjugate gradient method with line search [46].

Re-parametrization. The gradient-based learning proposed above has to be accomplished while respecting two sets of constraints: (i) the order constraints on ψ: $\{\psi_{j-1} \leq \psi_j \text{ for } j = 1, \ldots, L\}$, and (ii) the positive scale constraint on σ: $\{\sigma > 0\}$. To avoid constrained optimization, we introduce a re-parametrization of ψ using displacement variables δ_k, where $\psi_j = \psi_1 + \sum_{k=1}^{j-1} \delta_k^2$ for $j = 2, \ldots, L-1$. The positiveness constraint for σ is simply handled by introducing the free parameter σ_0 where $\sigma = \sigma_0^2$. Thus, the unconstrained parameters

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing

Algorithm 2: Copula Ordinal Regression Inference

Input: Parameter.: $\Omega = \{\varphi, \theta\}, G = \{E, V\}$
Data: $\mathcal{X} = \{x_i\}_{i=1}^{N}$
Result: Predictions: $\mathcal{Y}_p = \{y_i^p\}_{i=1}^{N}$

for $i : 1$ ***to*** N **do**
Compute Potentials
$\forall r \in V \rightarrow n_r = \phi_r(y_i^r|x_i, \varphi_r)$ (eq. 19)
$\forall (r,s) \in E \rightarrow e_{rs} = \phi_{rs}(y_i^r, y_i^s|x_i, \theta')$ (eq. 20)

if *Observations* y^o *are provided* **then**
Compute Conditional MAP solution (using LBP)
$y^* = \arg\min_y \Psi(y|y^o, V, E)$
else
Compute MAP solution (using LBP)
$y^* = \arg\min_y \Psi(y|V, E)$

of the ordinal marginals are $\{\beta, \psi_{\mathbf{1}}, \delta_{\mathbf{1}}, \ldots, \delta_{\mathbf{L-2}}, \sigma_{\mathbf{0}}\}$, and they are defined separately for each of the Q ordinal marginals, and stored in φ.

Training: During training, we seek to find optimal parameters Ω^* by solving the regularized optimization problem:

$$\Omega^* = \arg\min_{\Omega^*} NCL(\varphi, \theta) + \lambda_1 R_\varphi + \lambda_2 R_\theta \quad (16)$$

$$\lambda_1 R_\varphi = \sum_r \lambda_r R_{\varphi_r} = \sum_r \lambda_1^r \left\|\beta^{\mathbf{r}}\right\|^2 \quad (17)$$

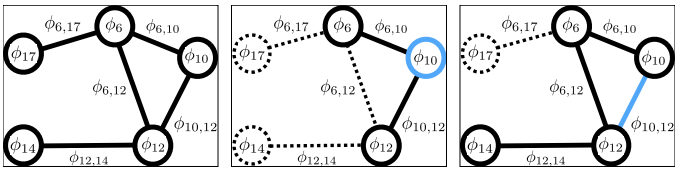
$$\lambda_2 R_\theta = \sum_{rs} \lambda_{rs} R_{\theta_{rs}} = \sum_{r,s} \lambda_2^{rs} \sqrt{\|w^{rs}\|^2 + \|\theta^{rs}\|^{-2}} \quad (18)$$

Where NCL is given by Eq.13, R_φ is the standard L_2 regularizer of the projection β and σ_0 of each unary. R_θ is the L_1 sparsity regularizer for each binary. The contribution of this regularization is two folded. First, the model converges rather to a sparse solution in which the binary weight parameter $w^{rs} = 0$. Secondly, the regularization ensures that the absolute value of θ is not zero, if $w \neq 0$. This makes the copula function stable and prevent the convergence to an independent model. In COR, λ_1^r and $\lambda_2^{r,s}$ are the regularization parameter and no specific regularization is necessary for the threshold parameters as they are automatically adjusted according to the score $\beta \mathbf{x}^\top$.

Solving for the parameters $\Omega = \{\varphi, \theta\}$ directly is possible, however, by noticing that the copula parameters θ are independent of the node potentials in the NCL, we can alternate between optimization of the marginals φ and the copula association θ. In this way, we detangle learning of the marginal model parameters from the joint copula parameters. Consequently, we reduce chances of falling into a local minimum due to the large number of parameters to be learned simultaneously. To this end, we propose a block-descent two-step optimization. We briefly describe the learning strategy.

The learning strategy is described in Alg. 1. Initially, we form an independence model by setting $E = \emptyset$ that treats each AU independently. After learning the parameters of the ordinal marginals $\{\varphi\}$, we consider a fully connected graph. We divide the parameter learning into two steps in which we optimize the θ and φ parameter independently.

θ-step: During the θ-step, we cycle through E and independently optimize the parameters of the bivariate copula function on a subset of E for each pair $(r,s) \in E$. This subsets is defined by the node (r,s) and the parent nodes of node r and node s. Note that this can be performed efficiently using parallel parameters estimation and the validation parameter θ_r and θ_{rs} can also be found independently for each subgraph using a validation set.



(a) Sparse Graph (b) φ_{10}-Step (c) $\theta_{10,12}$-Step

Fig. 3: Optimization steps of the COR learning algorithm. The dependencies are defined on (a) the sparse Graph. In (b) the φ step, the potential of the taget node (AU10) is conditionaly independent of all remaining nodes (AU17,AU14), given the adjacent nodes (AU6, AU12). The same rule is applied for in the θ (c) the edge potentials.

φ-step: Given the newly estimated copula parameters, in the φ-step, we minimize the objective function in Eq.16 w.r.t. the parameters of the ordinal marginals, i.e., φ. Specifically, we optimize the marginal parameters of each AU (φ^q) by using the unary and edge potentials where the target AU is present. We do so in parallel for all AUs. After the φ-step, we refine the association parameters θ.

We continue iterating between these two steps until convergence of the model performance on the validation set. We used the ICC(3,1) as for evaluation of the model performance during training. In our experiments, the algorithm converged in less than 5 iterations.

The advantage of the proposed learning approach over direct optimization is three-fold: (i) the estimation of the association and marginal parameters can be parallelized, thus leading to the computational complexity similar to that of marginal models. (ii) In the φ-step, we tune the regularization parameter λ separately for each AU and each Copula, using the balanced intensity levels for that AU (i.e., a subset of N training examples where the number of 0 intensity levels is balanced with the intensity 1). (iii) We can evaluate the model structure after each iteration and drop edges if either the weight parameter w or the association parameter θ converges to zero. This leads to a dynamic graph structure that can be efficiently learned. Note that in the case of the joint optimization, a single λ need be used, since cross-validation of AU-specific λ is infeasible. This process is summarized in Alg.1. ϕ_r is the conditional negative log likelihood of unary r and ϕ_{rs} the conditional log likelihood of binary rs. In all configurations of the COR framework, the potentials are defined as follows:

$$\phi_r(y_i^r|x_i) = \frac{\exp(s_r(y_i^r|x_i))}{\sum_{y'} \exp(s_r(y'|x_i))} \quad (19)$$

$$\phi_{rs}(y_i^r, y_i^s|x_i) = \frac{\exp(s_{rs}(y_i^r, y_i^s|x_i))}{\sum_{y_1', y_2'} \exp(s_{rs}(y_1', y_2'|x_i))} \quad (20)$$

Here, i is the frame number, s_r is the score function for node r, parameterized by φ_r and s_{rs} is the score function for edge rs, parameterized by θ_{rs}. These score functions are defined as follows:

$$s_r(y_i^r|x_i) = \ln(\Pr(y_i^r|x_i, \theta^r)) + \\ + \sum_{n \in par(r)} w_{rn} \ln(\Pr(y_i^r, |y_i^n, x_i, \varphi^{rn})) \quad (21)$$

(22)

$$s_{rs}(y_i^r, y_i^s|x_i) = w_{rs} \ln(\Pr(y_i^r, y_i^s|x_i, \theta^r)) + \\ + \sum_{m=r,s} \sum_{n \in par(m)} w_{nn} \ln(\Pr(y_i^n, |y_i^m, x_i, \varphi^{mn})) \quad (23)$$

One step of the optimization algorithm is depicted in Fig. 3. During the φ-step (3a) for AU10, only the model parameter that belong

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing

to its node and the parent nodes (AU6, AU12) and that from the corresponding edges are optimized. The regularization parameter are also tuned for the root node. The sames rule is applied during the θ step for the pair AU10-AU12. This approach has two advantages over the regular CRF optimization. First, the dependency of only directly connected AUs are modeled and higher order dependencies are ignored which speeds up the optimization. Secondly, the partition function can be computed per subgraphs, where each subgraph is defined by one root node and its parent nodes. In other words, we approximate the global joint dependency of all AUs but a product of local joint dependencies.

Inference: The inference of test data in undirected graphical models can be formulated as a discrete energy minimisation problem. This is in general np-hard due to the need to evaluate all possible label configurations. However, approximate methods based on Markov chain Monte Carlo (MCMC) and loopy belief propagation (LBP) for parameter learning have been proposed for making the parameter learning extremely efficient for subproblems [78]. We used the LBP. The running time for the LBP algorithm on our graph is expected to be $\mathcal{O}(M*N^k)$, where M is the number of AUs, N is the number of intensity levels for each AU, and k is the maximum clique size. Note that the algorithm gives exact marginals when the graph is a tree but only approximates the true marginals in loopy graphs. Fig. shows the experimental results of the model performance and fig. the inference time complexity. In our experiments, the best performance is reached in less than 100 iteration which makes the proposed model applicable for real-time application on an average work-station.

3.6 Partially Observed Data

The COR model, once trained, returns a multivariate probability distribution over the intensity levels of all AUs. By updating the inference algorithm, it is possible to make predictions on partially observed data, where some AU annotations of the test set are given. In this section, we briefly describe this approximate inference algorithm for partially observed outputs. The application for this could be as follows. A FACS coder can annotate a subset of AUs (AU6 and AU12, which are easy to identify) and the inference algorithm can then be used to compute the marginal node probability, conditioned on the AU intensity levels from the partially annotated subset. We define two sets of nodes. Y_p is the set of observed (annotated) AUs and Y_o is the set of unobserved AUs. The conditional probability of Y_p given Y_o is then defined by:

$$P(Y_p|Y_o) = \frac{P(Y_p, Y_o)}{P(Y_o)} \tag{24}$$

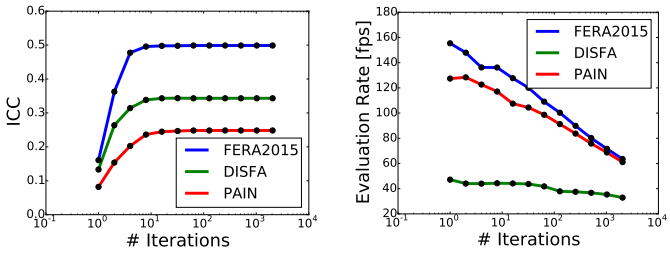
The marginal probability for the i-th AU having intensity k, given the labels Y_o can then be computed by marginalizing out the predictions for the remaining AUs from set Y_p.

$$P(Y_i = k|Y) = \sum_{Y_p \neq Y_i} P(Y_p|Y) \tag{25}$$

The marginal probably can be directly computed if the subset Y_p contains only a few AUs. In our experiments, we construct the graphical model as described in section 3.5. We fix the potentials that belong to AUs from Y_o to the states provided by the label but the potentials belonging to Y_p are still given by the potential-functions from the learned model. We apply LBP to approximate the node posteriors (see Alg.2 with partially observed data).

4 Experiments

In this section, we first describe the datasets and features and then show the comparisons with the state-of-the-art. In particular, we will show the results for the joint intensity estimation of AUs and individual AUs from the three datasets and compare its performance to that of the state-of-the-art methods. We also show the model performance on partially labeled data.



(a) classification perfomance (b) time complexity

Fig. 4: Complexity analysis of the LBP inference algorithms for different databases. As expected, the performance for the classification (4a) increases with the number of LBP iteration. It reaches its maximum in less than 100 iteration for all datasets. The evaluation rate in frames per second (4b) is linear decreasing with the number of LBP iteration.

4.1 Datasets and Experimental Procedure

Datasets

We evaluate the proposed model on major three benchmark datasets - UNBC-MacMaster Shoulder Pain Expression Archive (PAIN) [35], Denver Intensity of Spontaneous Facial Actions (DISFA) [39] and on that subset of the BinghamtonPittsburgh 4D Spontaneous Expression (FERA2015) [68] database that was used at the FERA2015 sub challenge for AU-intensity estimation. This databases include acted and spontaneous expressions and vary in image quality, video length, annotation, number of subjects, and context. The PAIN dataset contains video recordings of 25 patients suffering from chronic shoulder pain while performing a range of arm motion exercises. We applied 5-Fold cross validation (5 subjects per fold) on this dataset. The DISFA dataset contains video recordings of 27 subjects while watching YouTube videos. For this dataset we applied 9-Fold cross validation (3 subjects per fold). The FERA2015 database includes video of 41 participants (ages 18-29). There are 21 subjects in the training, 20 subjects in the development and 20 in the test partition. Since the test partition is not publicly available, we report our results on the development set. In all three datasets, each frame is coded in terms of the AU intensity on a six-point ordinal scale. For the experiments presented here, we used all 12 AUs from DISFA, all 10 AUs from PAIN, and from FERA 2015, we used all AUs that has been annotated with intensity levels (see the AU numbers and the distribution in that datasets in Fig. 5).

Features

We used the geometric facial features in our experiments, as in [22]. Namely, we used the locations of 49 out of 68 fiducial facial points provided for all from facial images in each dataset, We removed the points from the chin line, as these do not affect the estimation of target AUs. We then registered the 49 facial points to a reference face (average points in each dataset) using an affine transformation. To reduce the dimensionality of the features, we applied PCA, retaining 97% of the energy. This resulted in 20 dimensional feature vectors.

Evaluation metrics

In order to give a fair comparison for different tasks, we use the following measures for evaluation.

Mean Squared Error (MSE): MSE takes different scales into account and is commonly used to measure regression and ordinal classification performances [26], [47]. It also encodes how

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing

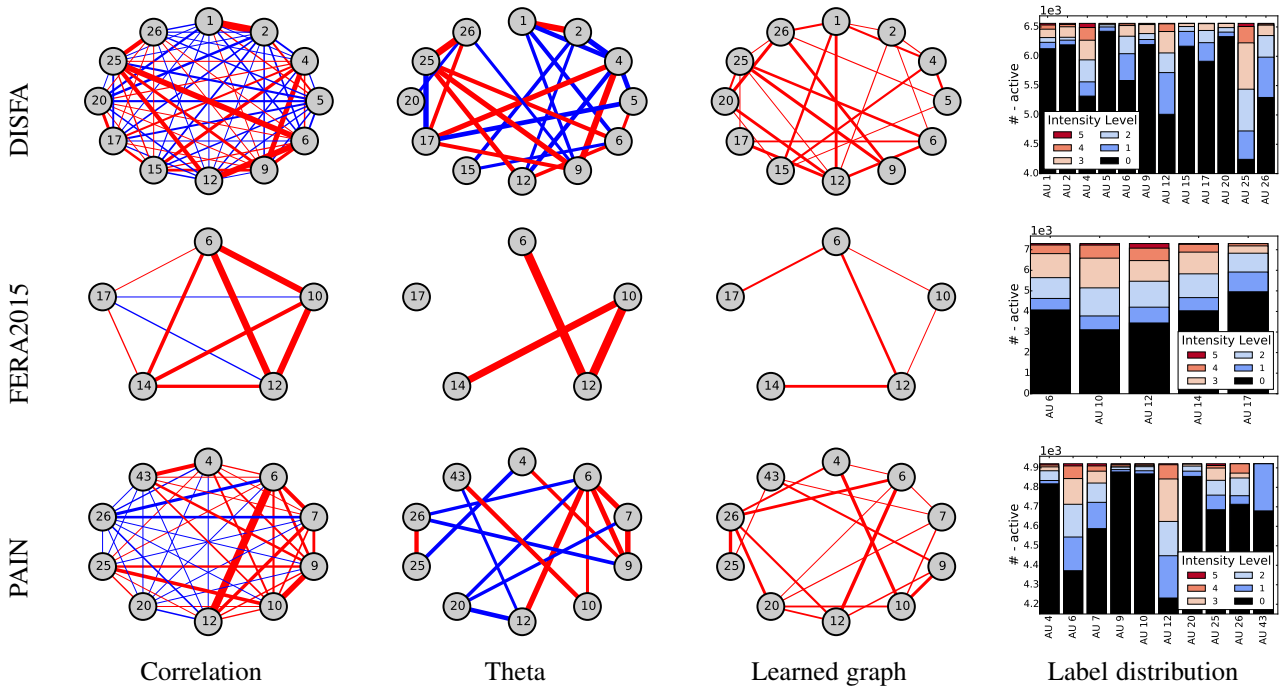


Fig. 5: The global AU relations depicted in terms of correlation coefficients. The negative corr is depicted in red, and positive corr in blue, while their magnitude is proportional to the thickness of the line. The learned association parameter $Theta$ and the $Weights$ show the learned graph for each database and the plots on the right show the product of this two parameters for each AU-pair which can be seen as edge importance for high performance.

inconsistent the classifier is in regard to the relative order of the classes, which is important when doing the intensity estimation.

Intra-class Correlation (ICC): We also report Intra-class Correlation (ICC(3,1) [56]), which is commonly used in behavioral sciences to measure agreement between annotators (in our case, the AU intensity labels and model predictions).

Models

We compare the performance of the proposed copula framework in three settings (COR-F, COR-IT and COR-HIT). We also compare our method to various state-of-the-art methods in the field.

Single-Output: As the baseline for the comparison, we use the results obtained by first applying the multi-class support vector machines (SVM) followed by the standard ordinal regression (SOR) [1]. Note that the SOR model uses the same marginal distribution functions as the node potentials of the proposed copula framework.

SVM and SOR were used as the baseline, by treating each of the intensity levels as a separate class. We optimized all hyperparameters by a grid search over the range $\left\{10^{\pm 4}, 10^{\pm 3}, ..., 0\right\}$ for the L2 regularizer, and selecting those that perform best on the validation set. We performed the parameter search for each AU separately.

Multi-Output: As the furthurer baseline for comparison, we also include the results attained by commonly used methods for classification and regression, i.e., multi-output K-Nearest-Neighbor (KNN), Multi-Layer-Perception (MLP), multivariate Gaussian Processes Regression (GP) and multivariate linear regression (MLR).

Finally, we compare our approach to the state-of-the-art methods for the robust AU intensity estimation - Latent Trees (LT-all) [22], MRF [50], our previous work on structured output [72], and the general CRF [28] for structured learning.

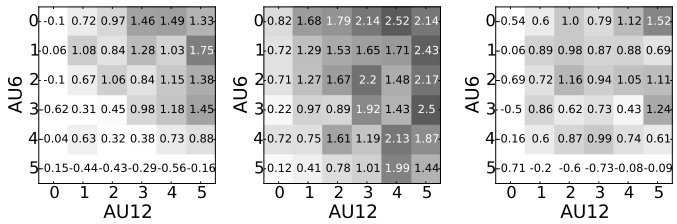
The authors of LT provided their source code and we implemented the MRF and CRF models. So the comparisons were performed in the same settings as our proposed models.

Deep-Models: We have also conducted a number of experiments using standard network architectures employed in previous works [13]. The basic **CNN [13]** model consists of a simple 2-layer networks with fully connected layer and softmax output-layer for multi-task classification. This model serves as a baseline for the deep models. Furthermore, we compare the proposed model with state-of-the-art networks for AU detect [80], and age estimation [43]. The **CNN-R [80]** was introduced for the task of AU detection and extends the basic **CNN** with a region layer. The weights of this layer are region specific, which makes the model adaptive and robust to background noise and illumination. We have trained this network from scratch for multi-task AU intensity estimation. The **CNN-O [43]** is a network with an ordinal classifier. It was introduced for the task of age estimation but can also be directly applied to AU intensity estimation of independent AUs. We also trained it from scratch for the target task. **VGG16 [57]** is a widely used very deep network for object detection. In order to adapt it for our task, we used the pre-trained model and fine-tuned the last 3 layer for AU intensity estimation. Lastly, the **SCNN [32]** is a deep structured network, introduced for mutli-task object detection. The linear pairwise potentials build a fully connected CRF, which are also trained using piecewise optimization [62]. In this work, the network is trained for multi-task AU intensity estimation, where each AU can be seen as a sepparate object.

4.2 Evaluation of the proposed models

Table 1 and table 4 show the performance for AU Intensity estimation on the FERA, DISFA and PAIN databases. As can be seen from Table 4. the single output ordinal model (SOR) performs consistently better than SVM. This is the case for 18 out of 27 AUs (in terms of ICC) and it is the outcome that the unconstrained SVM is more likely to overfit. The independent SOR model also achieves on average a better performance compared to the state-of-the-art LT method. Such performance of LT has also been observed by the authors of [22], who showed that significant improvements on highly noisy

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing



(a) DISFA (b) FERA2015 (c) PAIN

Fig. 6: Average association per intensity level of AU6&AU12 for all three databases. The association reaches the lowest value when one of the AUs is neutral (independent) and it is increasing with higher intensity.

features can be attained by the model due to its generative part. However, this robustness has not been observed in our experiments on the target data. MLR, MRF and the proposed COR frameworks consistently outperform the traditional models for joint AU intensity estimation. Clearly, these methods are all based on Bayesian networks and benefit from direct modelling of the dependencies between AUs trough edge potentials. This in contrast to the traditional methods (GP, MLP and MLR), that learn the AU dependency only indirectly (e.g. inner layer, latent states) but not on the decision level. Among the multi output methods, the worst performing method is the KNN. This is also expected since the KNN-classifer is highly affected by the curse of dimensionality. More precisely, the output space in this experiments has 6^Q possible configurations and there are simply no nearest neighbors with the same configuration to train the classifier. It is important to notice that the Bayesian models do not have this problem because they share the edge potentials only among a set of bivariate joints. Finally, all three COR models outperform the compared models by a large margin. We attribute this to their Bayesian networks properties that are combined with ordinal modelling of the marginals.

4.2.1 Evaluation of different Copula models

We discuss here the differences between COR models. The association parameter in the COR-IT model for a pair of AUs is shared over the full range of intensities. This results in suboptimal modelling of AU intensity pairs, especially with high intensity levels and on imbalanced data, where the majority of frames is not active. This effect can be observed on the DISFA and the PAIN dataset, where the COR-IT model is outperformed by the COR-HIT model. The COR-HIT model better fits the unbalanced datasets by learning different association for different pairs of intensities. This observation could not be made on the FERA2015 database. For this database, the sequences have been manually pre-segmented and only segments that show highly active facial expressions have been used in further processing. Fig. 5 shows that the distribution of intensity levels for all AUs in this database are better balanced than in PAIN or DISFA. However, such distribution of intensity levels is ideal but it can not be observed in naturalistic recordings. It is also important to notice that parts of the FERA2015 dataset (49 samples) were only partially annotated. This has a strong negative affect on the performance of all multi output models. We also notice that the joint inference by the proposed models consistently outperforms the compared multi-output models. This is also due to the individually tuned regularization parameters specifically for each AU, which is, in the LT, MRV, KNN and GP infeasible. Next, we observe that COR-HIT outperforms (on average) COR-IT across most of the AUs and in particular on the imbalanced databases (DISFA and PAIN) as expected. The COR-F model could not achieve such a good performance because the inference in this model is applied on the fully connected graph, which contains redundant edges. However, looking into ICC of AU9, we see that the COR-F performs significantly better for this AU. We attribute this to the full structure learning ability of the AU co-occurrence. This particular AU occurs very rarely but is correlated with AU10 and negative correlated with AU26 which occurs relatively frequently.

	Database:	FERA2015					
	AU:	6	10	12	14	17	avr.
ICC(3,1)	COR-HIT	**.76**	.72	**.81**	.29	.36	.59
	COR-IT	.74	**.73**	.79	.30	**.44**	**.60**
	COR-F	.72	.69	.79	.27	.40	.57
	MRF [50]	.72	.71	**.81**	**.33**	.30	.58
	LT-all [22]	.69	.58	.76	.30	.31	.53
	KNN [6]	.62	.57	.72	.11	.11	.43
	MLP [49]	.65	.67	.78	.27	.02	.48
	GP [75]	.68	.68	.78	.25	.23	.52
	CRF [28]	.67	.66	**.81**	.31	.28	.54
	MLR	.72	.70	.80	.27	.27	.55
	SOR	.68	.70	.79	.28	.26	.54
	SVM	.69	.65	.77	.18	.32	.52
MSE	COR-HIT	**0.97**	1.18	0.88	2.25	**0.81**	**1.22**
	COR-IT	1.04	1.20	0.92	2.04	0.89	**1.22**
	COR-F	1.17	1.19	1.04	2.48	0.93	1.36
	MRF [50]	1.06	1.18	0.94	2.00	1.17	1.27
	LT-all [22]	1.68	1.63	1.08	2.94	1.37	1.74
	KNN [6]	1.51	1.93	1.28	3.03	1.16	1.78
	MLP [49]	1.41	1.67	1.06	2.76	1.21	1.62
	GP [75]	1.25	1.17	0.95	**1.85**	0.96	1.24
	CRF [28]	1.16	1.26	**0.80**	1.89	1.15	1.25
	MLR	1.06	**1.16**	0.96	1.95	1.04	1.24
	SOR	1.33	1.40	1.06	2.75	1.25	1.56
	SVM	1.22	1.48	0.98	2.83	1.11	1.52

TABLE 1: Results on the FERA2015 (Train/Development) database.

4.2.2 Comparison to deep models

Table 5 shows the comparative results for the different deep models evaluated on the DISFA and PAIN datasets. These models are largely outperformed by traditional appearance based models (see Tab. 4). This is expected and was also observed in other studies and challenges (see [68]). However, the best performance for AU17 e.q. is achieved by the CNN-O model with ICC of 45%. This is 10% higher than the result of the best performing appearance based model (COR-HIT). AU17 couses the skin of the chin to wrinkle when activated. It is characterised exclusively by this textural change rather than by a change in facial morphology that we could capture as changes in facial landmark points locations. This important texture information is preserved in by the deep features but not by location of facial landmark, hence this result. Both, the CNN-O and the CNN-R model achieve an ICC of 29% on the DISFA dataset, which is the highest performance among the deep models. These models do not reach comparative results with our proposed model. The same applies for the CNN [13] and SCNN [32]. This confirms that to-data, standard benchmark datasets in the field of automatic facial coding do not provide sufficient data for deep models to be robustly trained on.

4.2.3 Evaluation on imbalanced data and head-pose variations

The proposed model consistently outperforms other models for all AUs for which the available data are strongly imbalanced. This is especially the case for AUs that are active in less than 8% of the frames (AU5, AU20 in Disfa and AU4, AU9, AU10 and AU20 in PAIN). The COR model achieves the highest performance on 5 out of 6 such AUs. We attribute this to the ability of (structured) learning the AU co-occurrence. In order to evaluate the robustness against of the proposed methods head pose variations, we removed all frontal images from the test data and evaluated the models on only that samples in which the orientation of the head to the camera is larger than 30

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing

Model	DISFA	PAIN
COR-HIT	**.33**	**.29**
COR-IT	.31	**.29**
COR-F	.29	**.29**
MRF [50]	.24	.23
LT-all [22]	.21	.23
KNN [6]	.11	.04
MLP [49]	.23	.17
GP [75]	.19	.19
CRF [28]	.26	.17
MLR	.12	.18
SOR	.23	.17
SVM	.18	.23

TABLE 2: ICC on non-frontal testdata

	Database:	FERA2015					
	AU:	6	10	12	14	17	avr.
ICC	COR-HIT	**.76**	**.72**	**.81**	.29	**.36**	**.59**
	MRF [50]	.72	.71	**.81**	**.33**	.30	.58
	CRF	.67	.66	**.81**	.31	.28	.54
+12	COR-HIT	**.84**	**.74**	1.00	**.46**	.38	**.68**
	MRF [50]	.80	.73	1.00	.34	.32	.64
	CRF	.76	.69	1.00	.33	**.39**	.63
+6	COR-HIT	1.00	**.77**	1.00	**.48**	.41	**.73**
	MRF [50]	1.00	.75	1.00	.30	.39	.69
	CRF	1.00	.69	1.00	.35	**.42**	.69
+10	COR-HIT	1.00	1.00	1.00	**.48**	.41	**.78**
	MRF [50]	1.00	1.00	1.00	.30	.41	.74
	CRF	1.00	1.00	1.00	.36	**.42**	.75

TABLE 3: Results on the development set from the FERA2015 database with partially observed data

degrees. By doing so, the performance in terms of ICC of the COR-HIT model dropped to 33% on the DISFA and to 29% on the PAIN dataset (see Table 4.2.3). However, among the compared methods, the best performing one is the CRF with an ICC of 26% on DISFA but only 17% on the PAIN dataset. We notice that the performance on the PAIN dataset is less effected if the frontal images are removed. This is also expected, since the recordings of that dataset were already non-frontal in most of the cases.

4.2.4 Evaluation on partially observed data

Table. 3 shows the results on the FERA2015 database with partially observed labels in the test data. We also include the results of the models based on Bayesian Networks for structured prediction. We first add the labels for AU12, which is the AU with the highest detection rate for all models. In the second iteration we add the labels for AU6 (second highest detection rate). Finally, the labels for AU10 are added in the third iteration. In all experiments, the performance for AU6 increases when labels for AU12 are provided. This is expected because of the strong dependency between this two AUs. This can also be observed for AU17 which has no dependency to other AUs and they to not co-occur. For example, by including the intensity levels of AU12 we can also infer some the states AU17 (being not active if AU12 is active). Therefore the increase in performance The COR model again outperforms the other models because it is capable to learn the sparse dependency structure directly from the training data. The MRF model is outperformed because its graph is restricted to have a tree like structure which is not the case in the COR-HIT model that can learn complex relationships among groups of AUs. Finally, the CRF model with unconstrained potential functions fails to predict partially observed data because it typically falls in a local minimum.

4.3 Qualitative Results

This section gives a more detailed evaluation of the models of the proposed copula framework on a single sequence of joy expression. As previously mentioned, The COR-HIT model is capable to model different association per intensity level. Typical AUs of joy are AU6 and AU12. Therefore we expect them to be highly positively correlated. More precisely, a high intensity of AU12 should increase the chance that AU6 is activated and vice versa. However, the AUs should be less dependent if they exhibit only a very low intensity. This heteroscadastic association, learned by the COR-HIT model, is show per Intensity pair in Fig.6. This association is close to 0 (independent association) if at least one AU exhibits a low intensity. The association increases with the AU intensity. Fig. 7 shows the predictions of AU6&12 of different models from the Copula framework and the related methods. Note that all models successfully predict the lower intensity levels, but they fail do detect the exact levels for the higher activations. However, this does not account for the COR-HIT model. Due to its heteroscadastic adaptation, it is able to predict exactly all intensity levels, the lower and the higher ones. Finally, Fig. 2 shows the average negative log likelihood (NCLL) (goodness of fit) for the Independent, the Frank, the Gumbel, and the Clyton copula on all three databases. The lowest NCLL (the best fit) is reached by the Frank Copula. Gumbel and Clyton might be a good choice for other problems but they fail to model the negative correlations. Another important property of the Franc copula is its ability to model independence by setting the association parameter to a very low number. In our experiments, we found that for values $|\theta| < 10^{-5}$ the Frank copula acts exactly like the independent copula and the edge potentials of the COR model are independent. In this special case, the COR model results in the standard ordinal regression model.

5 Conclusions

In this paper, we proposed a novel Copula Ordinal Regression Framework for joint modeling and estimation of intensities of multiple AUs from facial images. The proposed model was evaluated in three different settings and it has been shown experimentally that it significantly improves the intensity estimation performance by modelling the structured ordinal output. First, we showed that by endowing the model with separate but coupled marginal and dependency components, we can successfully capture correlations between different facial features and co-occurrences of various AUs. This approach generalizes prior methods that rely on independent models by using an efficient parametric and flexible representation of the copula functions tied together through a CRF model. Secondly, we demonstrate that the proposed Copula Framework outperforms related independent models and the state-of-the-art approaches for joint intensity estimation of AUs. We demonstrate on three different datasets that the heteroscadastic COR-HIT model predicts the AU intensities the best, particularly in the case of imbalanced data with strong head-pose variations. Lastly, we extend the inference algorithm for prediction of partially observed test data and show that the proposed model outperforms related models that are based on Bayesian networks for structured prediction.

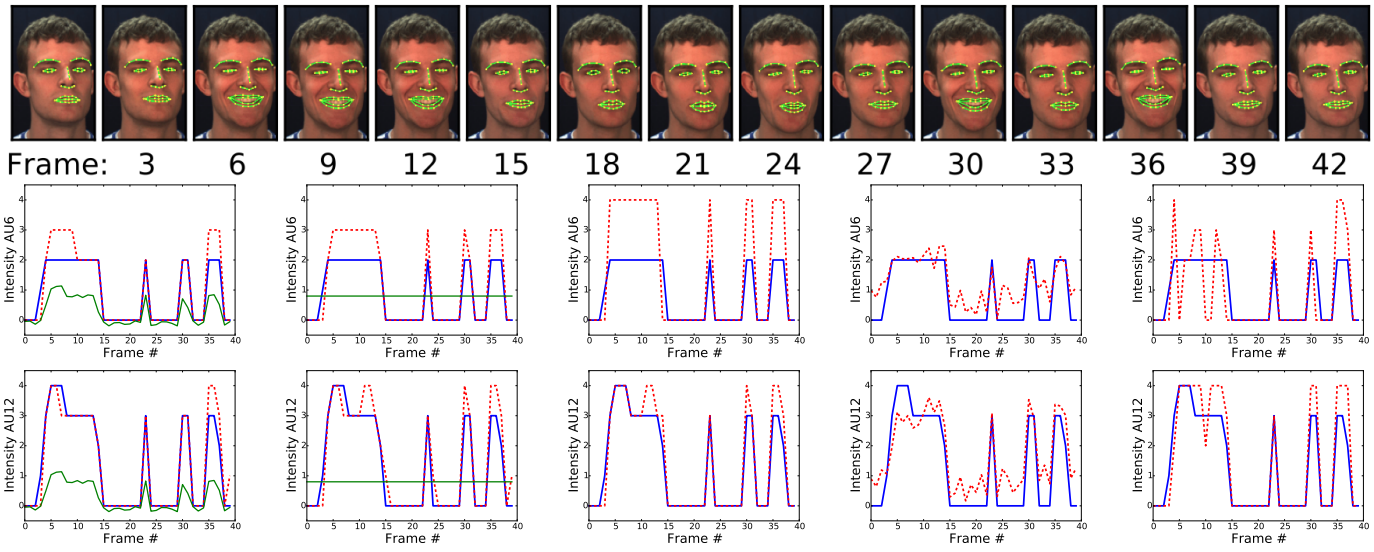
Acknowledgments

This work has been funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA), and no. 688835 (DE-ENIGMA). The work of O. Rudovic is also funded by European Union H2020, Marie Curie Action - Individual Fellowship (EngageMe 701236). The work of Vladimir Pavlovic has been funded by the National Science Foundation under Grant no. IIS0916812.

References

[1] A. Agresti. Analysis of ordinal categorical data. *Wiley Series in Prob. and Stat.*, pages 1–287, 1984.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2728534, IEEE Transactions on Affective Computing



(a) COR-HIT (b) COR-IT [71] (c) SOR (d) GP [75] (e) SVM

Fig. 7: Intensity estimation of AU6&AU12 from the DISFA database attained by COR-HIT, COR-IT, SOR, GP and SVM. The upper figure shows the series of input images with landmarks and the corresponding AU intensity levels. The lower figures show the true (solid blue) and predicted (dashed red) labels by the models. The copula models also show the value of the association parameter (green) for the target pair (AU6&AU12). For clarity, we have downscaled the value of the association parameter by 10.

	Database:	DISFA													PAIN										
	AU:	1	2	4	5	6	9	12	15	17	20	25	26	avr.	4	6	7	9	10	12	20	25	26	43	avr.
ICC(3,1)	COR-HIT	.53	**.48**	.63	**.36**	.47	.37	.8	**.35**	.35	.12	.84	**.56**	**.49**	**.1**	**.53**	**.38**	.39	.61	**.5**	.0	**.38**	.06	.19	**.31**
	COR-IT	**.55**	.46	.65	.32	.38	.42	.78	.31	**.39**	.15	.84	.54	.48	.07	.49	.36	.39	.6	**.5**	.01	.36	.03	**.22**	.3
	COR-F	.46	.41	.53	.23	.39	**.47**	.77	.19	.23	.12	.83	.52	.43	.02	.51	.34	**.42**	**.63**	.48	.0	.32	.1	.13	.3
	MRF [50]	.48	.37	.6	.31	.35	.3	.78	.27	.25	.06	.85	.47	.43	.01	.43	.36	.04	.6	.43	.0	.33	.0	.13	.23
	LT-all [22]	.37	.31	.45	.13	.52	.29	.77	.15	.3	.05	.78	.49	.38	.01	.41	.33	.33	.61	.44	.0	.33	.0	.15	.26
	KNN [6]	.28	.26	.34	.09	.36	.22	.7	.12	.16	.1	.78	.31	.31	.03	.27	.08	.0	.02	.26	.0	.29	.03	.1	.11
	MLP [49]	.31	.12	.6	.05	.49	.36	.73	.1	.24	.01	.77	.49	.35	.01	.45	.21	.03	.43	.42	.0	.34	**.12**	.12	.21
	GP [75]	.12	.1	**.68**	.0	**.58**	.43	.83	.01	.27	.0	.84	.53	.36	.07	.43	.26	.1	.19	.41	.01	.29	.08	.19	.2
	CRF [28]	.41	.36	.46	.16	.52	.29	.78	.16	.31	.08	.78	.50	.40	.06	.45	.27	.03	.12	.43	.00	.23	.07	.01	.17
	DSRVM [23]	.26	.22	.47	.14	.47	.40	.75	.28	.34	**.19**	.58	.35	.37	-	-	-	-	-	-	-	-	-	-	.21
	MLR	.04	.08	.64	.0	.5	.35	.81	.0	.15	.0	.85	.46	.32	.04	.45	.17	.17	.29	.43	**.07**	.33	.07	.13	.22
	SOR	.34	.28	.51	.12	.46	.31	**.82**	.16	.26	.1	**.86**	.51	.39	.04	.39	.21	.23	.43	.37	.0	.32	**.12**	.12	.22
	SVM	.25	.19	.45	.12	.53	.25	.78	.13	.2	.05	.77	.42	.34	.07	.46	.22	.15	.42	.45	.03	**.38**	.01	.09	.23
MSE	COR-HIT	.38	**.34**	.82	**.06**	.41	**.25**	**.28**	**.14**	**.27**	**.14**	**.43**	**.37**	**.32**	**.14**	**.44**	**.29**	**.05**	**.03**	**.54**	.09	.26	.28	**.04**	**.22**
	COR-IT	**.36**	.35	.83	.07	.52	.29	.3	.15	.3	.18	.47	.39	.33	.15	.46	**.29**	**.05**	.04	**.54**	.08	.27	.29	**.04**	**.22**
	COR-F	.4	.37	.97	.08	.54	.25	.37	.17	.34	.19	.54	.41	.39	**.14**	.47	.31	.06	**.03**	.6	.07	.28	.29	.05	.23
	MRF [50]	.41	.38	.9	.08	.62	.31	.39	.15	.33	.18	.49	.5	.4	.15	.58	.3	.06	.04	.68	.09	.28	.3	.05	.25
	LT-all [22]	.4	.36	.91	**.06**	.37	.28	.36	.15	.28	.15	.62	.41	.36	.15	.57	.31	.06	.04	.68	.06	.29	.29	.05	.25
	KNN [6]	.57	.52	1.43	.11	.6	.38	.54	.2	.39	.22	.72	.65	.53	.18	.5	.4	.08	.06	.64	.06	.28	.34	.05	.26
	MLP [49]	.42	.42	.74	.23	.37	.28	.41	.15	**.27**	.17	.63	.41	.37	.15	.55	.36	.06	.04	.72	**.05**	.28	.32	.05	.26
	GP [75]	.59	.51	.72	.17	.43	.36	.38	.26	.36	.25	.56	.49	.42	.24	.54	.39	.16	.16	.66	.15	.33	.37	.14	.31
	CRF [28]	.39	.36	.93	.07	.39	.28	.36	.15	.29	.15	.63	.41	.37	.15	.49	.32	.07	.06	.62	.06	**.25**	**.27**	.05	.23
	MLR	.54	.43	**.8**	.07	.44	.32	.34	.17	.32	.15	.53	.5	.39	.19	.82	.48	.09	.07	1.02	.08	.35	.37	.06	.35
	SOR	**.36**	**.34**	.79	.12	.48	.3	.34	**.14**	.3	.2	.53	.4	.36	.15	.58	.35	.11	.05	.74	**.05**	.3	.3	.05	.27
	SVM	.44	.39	.91	.07	**.34**	.28	.34	**.14**	.28	.15	.63	.41	.37	.16	.55	.34	.07	.04	.67	.06	.31	.33	.07	.26

TABLE 4: Average performance of the apperance based models tested on the DISFA and Shoulder-Pain database for intensity estimation. Note that the models from the Copula framework outperform the traditional models and the state-of-the-art in 15 out of 22 AUs and reaches the overall best average performance in terms of ICC and MSE.

	Database:	DISFA													PAIN										
	AU:	1	2	4	5	6	9	12	15	17	20	25	26	avr.	4	6	7	9	10	12	20	25	26	43	avr.
ICC(3,1)	CNN [13]	.05	.04	.36	.02	**.44**	.27	.67	.25	.08	.03	.46	.22	.23	.11	.50	.23	.25	**.28**	**.57**	.08	.43	**.16**	.23	**.29**
	CNN-R [80]	.05	.06	.32	.02	.36	**.39**	**.77**	**.29**	.19	.04	**.65**	.35	**.29**	.11	**.51**	**.28**	.25	.22	.54	.07	**.46**	.15	**.24**	.28
	CNN-O [43]	.04	.05	**.41**	.01	.35	.19	.72	.23	**.45**	**.06**	.53	**.44**	**.29**	**.21**	.45	.13	.07	.22	.52	.07	.18	.09	**.24**	.21
	SCNN [32]	.03	.07	.01	.00	.29	.08	.67	.13	.27	.00	.59	.33	.20	.11	.40	.28	.25	.22	.41	.16	.26	.11	.21	.24
	VGG16 [57]	**.19**	**.14**	.19	.02	.39	.33	.68	.14	.27	.03	.59	.38	.28	.11	**.51**	**.28**	**.30**	.22	.54	**.10**	**.46**	.15	**.24**	**.29**

TABLE 5: Average performance of the models based on CNN's. Note that the models from the Copula framework (Tab. 4) outperform the CNN based models by a large margin, Especially on the DISFA dataset.

[2] Z. Ambadar, J. F. Cohn, and L. I. Reed. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of nonverbal behavior*, 33(1):17–34, 2009.

[3] T. Baltrusaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *FG*, volume 6, pages 1–6. IEEE, 2015.

[4] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006.

[5] P. Berkes, F. Wood, and J. W. Pillow. Characterizing neural dependencies with copula models. In *NIPS*, pages 129–136. 2009.

[6] T.-H. Chiang, H.-Y. Lo, and S.-D. Lin. A ranking-based knn approach for multi-label classification. *ACML*, 25:81–96, 2012.

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

[8] P. Ekman, W. V. Friesen, and J. C. Hager. Facial action coding system. *Manual: A Human Face*, 2002.

[9] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *ICCV*, 2015.

[10] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, pages 432–441, 2008.

[11] C. Genest. Frank's family of bivariate distributions. *Biometrika*, pages 549–555, 1987.

[12] C. Genest and J. Nešlehová. A primer on copulas for count data. *Astin Bulletin*, 37(02):475–515, 2007.

[13] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *FG*, volume 6, pages 1–5. IEEE, 2015.

[14] R. Herbrich, T. Graepel, and K. Obermayer. Support vector learning for ordinal regression. In *NIPS*, volume 1, pages 97–102. IET, 1999.

[15] S. Horvath. Weighted network analysis: Applications in genomics and systems biology. *Springer Science and Business Media*, 2011.

[16] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):765–781, 2011.

[17] S. Jaiswal, B. Martinez, and M. F. Valstar. Learning to combine local models for facial action unit detection. In *FG*, volume 6, pages 1–6. IEEE, 2015.

[18] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De La Torre. Continuous au intensity estimation using localized, sparse facial feature space. In *FG*, pages 1–7. IEEE, 2013.

[19] B. Jiang, B. Martinez, and M. Pantic. Parametric temporal alignment for the detection of facial action temporal segments. In *British Machine Vision Conference*, Nottingham, UK, September 2014.

[20] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modelling. *IEEE Transactions on Cybernetics*, 44(2):161–174, 2014.

[21] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. *International Symposium on Visual Computing*, pages 368–377, 2012.

[22] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, 2015.

[23] S. Kaltwang, S. Todorovic, and M. Pantic. Doubly sparse relevance vector machine for continuous facial behavior estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1748–1761, September 2016.

[24] A. Kapoor and R. W. Picard. Multimodal affect recognition in learning environments. In *ACM*, pages 677–682. ACM, 2005.

[25] P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[26] M. Kim and V. Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. *ECCV*, pages 649–662, 2010.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] A. Kulesza and F. Pereira. Structured learning with approximate inference. *NIPS*, pages 785–792, 2007.

[29] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[30] 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*, pages 1–7, 2013.

[31] Y. Li, B. Wu, B. Ghanem, Y. Zhao, H. Yao, and Q. Ji. Facial action unit recognition under incomplete data based on multi-label learning with missing labels. *Pattern Recognition*, 60:890–900, 2016.

[32] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[33] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *FG*, pages 1–6, 2013.

[34] 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. *TSMCB*, pages 664–674, 2011.

[35] 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 *FG*, pages 57–64, 2011.

[36] S. Lucey, A. B. Ashraf, and J. F. Cohn. *Investigating spontaneous facial action recognition through aam representations of the face*. INTECH Open Access Publisher, 2007.

[37] M. Mahoor, S. Cadavid, D. Messinger, and J. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. *CVPR*, pages 74–80, 2009.

[38] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. In *FG*, pages 336–342, 2011.

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

[40] R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *JMLR*, pages 781–794, 2012.

[41] Z. Ming, A. Bugeau, J.-L. Rouas, and T. Shochi. Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine. In *FG*, volume 6, pages 1–6. IEEE, 2015.

[42] J. Nicolle, K. Bailly, and M. Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *FG*, volume 6, pages 1–6. IEEE, 2015.

[43] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, June 2016.

[44] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[45] J. E. Pessa, V. P. Zadoo, P. A. Garza, E. K. Adrian, A. I. Dewitt, and J. R. Garza. Double or bifid zygomaticus major muscle: anatomy, incidence, and clinical correlation. *Clinical Anatomy*, pages 310–313, 1998.

[46] C. Rasmussen and C. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

[47] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *TPAMI*, pages 944–958, 2015.

[48] A. Ruiz, J. Van de Weijer, and X. Binefa. From emotions to action units with hidden and semi-hidden-task learning. In *ICCV*, pages 3703–3711, 2015.

[49] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[50] G. Sandbach, S. Zafeiriou, and M. Pantic. Markov random field structures for facial action unit intensity estimation. In *ICCV*, 2013.

[51] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. A dynamic approach to the recognition of 3d facial expressions and their temporal models. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), Special Session: 3D Facial Behavior Analysis and Understanding*, pages 406–413, Santa Barbara, CA, USA, March 2011.

[52] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012. 3D Facial Behaviour Analysis and Understanding.

[53] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2015.

[54] A. Savran, B. Sankur, and M. T. Bilge. Regression-based intensity estimation of facial action units. *IMAVIS*, 30(10):774–784, 2012.

[55] J. H. Shih and T. A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, pages 1384–1399, 1995.

[56] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[57] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[58] A. Sklar. Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture notes-monograph series*, pages 1–14, 1996.

[59] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor. Exploiting sparsity and co-occurrence structure for action unit recognition. In *FG*, volume 1, pages 1–8. IEEE, 2015.

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

[61] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[62] C. Sutton and A. McCallum. Piecewise pseudolikelihood for efficient training of conditional random fields. In *Proceedings of the 24th international conference on Machine learning*, pages 863–870. ACM, 2007.

[63] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems*, pages 2553–2561, 2013.

[64] S. Taheri, Q. Qiu, and R. Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *TIPS*, 23(8):3590–3603, 2014.

[65] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):258–273, Feb 2010.

[66] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *TPAMI*, pages 1683–1699, 2007.

[67] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *IJDWM*, pages 1–13, 2007.

[68] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *FG*, volume 6, pages 1–8. IEEE, 2015.

[69] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analyis of the first facial expression recognition challenge. *IEEE Transactions of Systems, Man and Cybernetics – Part B*, 42(4):966–979, 2012.

[70] M. F. Valstar, S. Zafeiriou, and M. Pantic. *Facial action recognition in 2D and 3D*, pages 167–186. IGI Global, Hershey, PA, USA, 2014.

[71] R. Walecki, O. Rudovic, M. Pantic, and V. Pavlovic. Copula ordinal regression for joint estimation of facial action unit intensity. *CVPR*, June 2016.

[72] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic. Variable-state latent conditional random fields for facial expression recognition and action unit detection. In *FG*, pages 1–8, Ljubljana, Slovenia, May 2015.

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

[74] R. Winkelmann. *Econometric analysis of count data*. Springer Science & Business Media, 2003.

[75] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. *ICML*, pages 1012–1019, 2005.

[76] A. Yuce, H. Gao, and J.-P. Thiran. Discriminant multi-label manifold embedding for facial action unit detection. In *FG*, volume 6, pages 1–6. IEEE, 2015.

[77] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *TKDE*, pages 1338–1351, 2006.

[78] Y. Zhang and J. Schneider. A composite likelihood view for multi-label classification. In *AISTATS*, pages 1407–1415, 2012.

[79] 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 *CVPR*, 2015.

[80] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

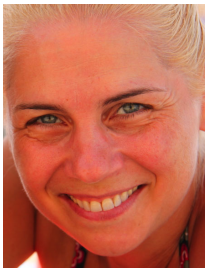
[81] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, pages 2562–2569. IEEE, 2012.

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

Robert Walecki received his MSc degree in Physics from the The Ruprecht-Karls-University in Heidelberg, Germany, in 2013. He is currently working towards his Ph.D. degree at the Department of Computing, Imperial College London, London, UK. His research interests span the areas of Computer Vision, Pattern Recognition, Machine Learning and, in particular, human-computer interaction and automatic human behavior analysis.

Ognjen Rudovic received his BSc degree in Automatic Control from Faculty of Electrical Engineering, University of Belgrade, Serbia, in 2007, MSc degree in Computer Vision from Computer Vision Center (CVC), Universitat Autonoma de Barcelona, Spain, in 2008, and PhD in Computer Science from Imperial College London, UK. He is currently a Research Fellow at the Computing Department, Imperial College London, UK. His research interests are in automatic recognition of human affect, machine learning and computer vision.

Vladimir Pavlovic received the PhD degree in electrical engineering from the University of Illinois at Urbana-Champaign in 1999. From 1999 until 2001, he was a member of the research staff at the Cambridge Research Laboratory, Massachusetts. He is an associate professor in the Computer Science Department at Rutgers University, New Jersey. Before joining Rutgers in 2002, he held a research professor position in the Bioinformatics Program at Boston University. His research interests include probabilistic system modeling, time-series analysis, computer vision, and bioinformatics.



Maja Pantic is Professor in Affective and Behavioural Computing at Imperial College London, Computing Dept., UK, and at the University of Twente, Dept. of Computer Science, Netherlands. She received various awards for her work on automatic analysis of human behaviour including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She currently serves as the Editor in Chief of Image and Vision Computing Journal, and as an Associate Editor for IEEE Trans. on Systems, Man, and Cybernetics Part B and IEEE TPAMI.