Context-sensitive Dynamic Ordinal Regression for Intensity Estimation of Facial Action Units

Ognjen Rudovic, *Student Member, IEEE*, Vladimir Pavlovic, *Senior Member, IEEE* and Maja Pantic, *Fellow, IEEE*

Abstract—Modeling intensity of facial action units from spontaneously displayed facial expressions is challenging mainly because of high variability in subject-specific facial expressiveness, head-movements, illumination changes, etc. These factors make the target problem highly context-sensitive. However, existing methods usually ignore this context-sensitivity of the target problem. We propose a novel Conditional Ordinal Random Field (CORF) model for context-sensitive modeling of the facial action unit intensity, where the W5+ (*who*, *when*, *what*, *where*, *why* and *how*) definition of the context is used. While the proposed model is general enough to handle all six context questions, in this paper we focus on the context questions: *who* (the observed subject), *how* (the changes in facial expressions), and *when* (the timing of facial expressions and their intensity). The context questions *who* and *how* are modeled by means of the newly introduced context-dependent covariate effects, and the context question *when* is modeled in terms of temporal correlation between the ordinal outputs, i.e., intensity levels of action units. We also introduce a weighted softmax-margin learning of CRFs from data with skewed distribution of the intensity levels, which is commonly encountered in spontaneous facial data. The proposed model is evaluated on intensity estimation of pain and facial action units using two recently published datasets (UNBC Shoulder Pain and DISFA) of spontaneously displayed facial expressions. Our experiments show that the proposed model performs significantly better on the target tasks compared to the state-of-the-art approaches. Furthermore, compared to traditional learning of CRFs, we show that the proposed weighted learning results in more robust parameter estimation from the imbalanced intensity data.

Index Terms—FACS, action unit intensity, spontaneous facial behavior, facial expression analysis, ordinal regression, conditional random fields, context modeling

1 Introduction

Faces hold valuable clues to people's emotions and intentions. Facial expressions are some of the most direct, naturally preeminent means for human beings to regulate interactions with each other [1]. They communicate emotions, clarify and stress what is being said, and signal comprehension, disagreement and stances. Machine understanding of facial expressions could revolutionize user interfaces for artifacts such as robots, mobile devices, cars, and conversational agents [2]. Other valuable applications are in the domain of medicine and psychology, where it can be used to improve medical assistance as well as develop automated tools for behavioral research. Therefore, machine understanding of facial expressions has recently become a hot research topic.

Facial expressions are usually described in terms of variation in configuration and strength of facial muscle actions. To this aim, the Facial Action Coding System (FACS) [3] defines a comprehensive set of atomic non-overlapping facial muscle actions named Action Units (AUs) [4]. Each and every facial expression can be described in terms of these AUs and their intensities. Specifically, FACS defines 9 different AUs in the upper face, 18 in the lower face, and 5 AUs that cannot be classified as belonging to either the upper or the lower face. It also defines the so-called action descriptors (ADs), 11 for head position, 9 for eye position, and 14 additional descriptors for miscellaneous actions. FACS also provides the rules for scoring the intensity of each AU in a 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 ADs, AUs and their intensity that produced the expression. However, this process is tedious and error-prone due to the large number of AUs and the difficulty in discerning their intensities.

To date, most of the work on automated analysis of AUs has focused on detection of the presence/absence of AUs (e.g. [5], [2], [6], [7], [8]) instead of their full range intensity estimation. Yet, the meaning and function of spontaneous facial expressions depends largely on intensity of AUs. For example, the smiles of enjoyment are full-blown smiles, while the "fake happiness smiles" (as in sarcasm) may be asymmetric and are usually of lower intensity when observed in naturalistic social settings. As noted in [9], "most of the smile genuineness impression is created by the intensity [and the facial motion] of the smile, not just the activation of AU6". However, discerning different intensities of AUs is a far more challenging task than AU detection for several reasons. First and foremost, the perceived intensities of AUs depend greatly on the facial morphology and expressiveness of the observed subject. As noted in studies on human anatomy (e.g., [10]) as well as

- *O. Rudovic is with the Dept. of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK. E-mail: o.rudovic@imperial.ac.uk,*
- *V. Pavlovic is with the Dept. of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019.E-mail: vladimir@cs.rutgers.edu*
- *M. Pantic is with the Dept. of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK, and with the Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, The Netherlands. E-mail: m.pantic@imperial.ac.uk*

Manuscript received.

in the FACS manual, "the intense muscular contractions are combined with the individual's physical characteristics to produce changes in appearance that then vary somewhat between different subjects". Also, each subject may have a different aptitude for expressivity (e.g., extrovert vs. introvert people). This, in turn, makes it difficult to grasp what constitutes the maximal level of appearance change for each subject. For example, different people gesticulate differently and while some usually display very broad smiles, others display small, less-wide smiles. This, in particular, may be due to the high inter-personal variability in the morphology of the zygomatic major muscle [10], the activity of which results in a smile. Second, co-occurrences of AUs affect the criteria for scoring their intensity. For example, 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 [3]. Third, a change in lighting, head position, and transient shadows can all give the impression of a different AU intensity. All these factors make the AU intensity estimation a very challenging task, and, above all, highly context-sensitive. Hence, to determine the intensity of AUs accurately, one must also know the context in which they occur [11].

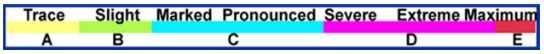


Fig. 1: Relationship between the scale of facial appearance change and intensity levels when evidence of an AU is present [3].

To this end, we propose a Context-sensitive Conditional Ordinal Random Field (cs-CORF) model for dynamic estimation of AU intensity levels. This model is based on the linear-chain CRF model [12] for sequence classification. To impose ordering constraints on AU intensity levels, we define the node features in the model using the modeling framework of (static) ordinal regression models [13], [14]. More importantly, we extend this framework by accounting for the omnipresent impact of context on AU intensity estimation via modeling of context-sensitive variability in data. To this end, we adopt the widely accepted *W5+* context model [11], where the following six questions are used to summarize the key aspects of the context in which the target action (in our case, the AU intensity) occurs: *who* (the subject's identity, age and expressiveness), *where* (environmental characteristics such as illumination), *what* (task-related cues of the facial action such as head tilts, nods, etc.), *how* (the information is passed on by means of facial expression intensity), *when* (timing of facial expressions and their intensity) and *why* (the context stimulus such as the humorous videos). Existing approaches to AU intensity estimation (e.g., [4], [15], [16]) are context-free as they model the context question *how* only, without taking into account the other context questions. By contrast, the proposed cs-CORF model can be used to model all six context questions. We demonstrate this on the context questions *who*, *how* and *when*; however, the other context questions can be modeled in a similar manner. The context questions *who* and *how* are accounted for at the feature level by newly introduced Context-related Covariate Effects (CRE) and Context-free Covariate Effects (CFE), where the CFE coincide with those modeled in the context-free models. These two effects are efficiently embedded via *ordinal* node potentials of the cs-CORF model. On the other hand, the context question *when* is addressed at the model level by encoding temporal dependence between the intensity labels via the edge potentials of the model. We do this at the model level in order to avoid the potential problem of temporal misalignment of raw image features. The CRE component is considered constant along the sequence, and is derived from the subjects' characteristics such as their facial shapes (when there is no AU activation present). This component is of particular importance as it directly accounts for the subject-specific bias in the model parameters. We also account for heterogeneity of subjects by modeling heteroscedasticity of both the CRE and CFE components. This allows the model to further capture the expressiveness of each subject. All these effects are summarized in the graphical representation of the proposed cs-CORF model shown in Fig. 2. Lastly, to address the problem of label/level imbalance in a principled manner, we introduce a weighted softmax-margin learning approach for CRFs, based on a generalization of the slack and margin rescaling modeling criteria in [17], [18].

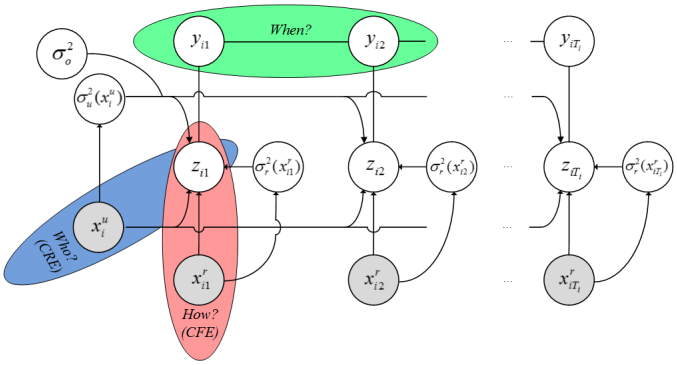


Fig. 2: **The proposed cs-CORF model.** The input to the model are the time-varying CFE covariates (x^r_{ij}) and the constant (on the sequence level) CRE covariates (x^u_i), used to model the context-questions *how* and *who*, respectively. These effects are linearly related to the latent variable z_i, contaminated by Gaussian noise with zero mean and variance defined as the sum of the CRE ($\sigma^2_u(x^u_i)$) and CFE ($\sigma^2_r(x^r_{ii})$) heteroscedastic variance, as well as σ^2_o that accounts for unexplained variance in the data. The latent variable z_i is non-linearly mapped to the ordinal labels y_i via the probit link function, used to define the node potentials of the cs-CORF model. The context question *when* is modeled by encoding the first-order temporal dependences between the intensity levels via the edge potentials of the model.

We demonstrate performance of the proposed model on the UNBC Shoulder Pain [19] and DISFA [20] datasets of spontaneously displayed facial expressions. Compared to existing approaches for the target task, we show that there is a significant increase in the AU/facial expression intensity estimation performance when the proposed cs-CORF model is used. We also show that the proposed weighted softmax-margin learning approach results in a more robust parameter learning compared to the standard *maximum a posteriori* (MAP) estimation of CRFs.

2 Related Work

The automated estimation of AU intensity is a relatively recent problem within the field, and only a few works have addressed it so far. These can be divided into classification-based methods [4], [20] and regression-based methods [16], [21], [22]. The former use classifiers for nominal data, such as the Support Vector Machine (SVM), to classify the intensity levels of AUs. Specifically, Mahoor et al. [4] addressed the intensity estimation of AU6 (cheek raiser) and AU12 (lip corner puller) from facial images of infants. Input features were obtained by concatenation of the facial shape (facial landmarks) and appearance (gray-level intensity of each pixel), which were pose-normalized using an Active Appearance Model (AAM) [19]. Due to the excessive number of features, the Spectral Regression (SR) [23] was applied to select the most relevant for each AU. The intensity classification based on these features was then performed by the SVM. Mavadati et al. [20] proposed a new dataset of spontaneously displayed facial expressions, named DISFA, and employed the same approach as in [4] for intensity classification of 13 AUs. The evaluation was carried out on different sets of features derived from the facial appearance using Local Binary Patterns (LBPs), histograms of oriented gradients (HOGs) and Localized Gabor Filters (LGF).

The regression-based methods model the AU intensity on a continuous scale using the logistic-regression-based models [16], Relevance Vector Machine (RVM) regression [21], and Support Vector Regression (SVR) [22]. For instance, Savran et al. [16] proposed a model based on logistic regression for AU intensity estimation. The model was evaluated on the Bosphorus Database [24], which contains 3D facial images of posed facial expressions coded in terms of 25 AUs and their intensities. To select input features, the authors applied an AdaBoost-based method to Gabor wavelet magnitudes of 2D luminance and 3D geometry extracted from the target images. Kaltwang et al. [21] used the RVM model for intensity estimation of spontaneously displayed facial expressions of pain and 11 AUs from the Shoulder-pain dataset [19]. The effectiveness of different image features such as Local Binary Patterns (LBPs), Discrete Cosine Transform (DCT) and facial landmarks, as well as their fusion, was evaluated for the target task. Jeni et al. [22] proposed a sparse representation of the facial appearance obtained by applying Non-negative Matrix Factorization (NMF) filters to gray-scale image patches extracted around facial landmarks. The image patches were then processed by applying personal mean texture normalization, and used as input to the SVR. The model was evaluated on intensity estimation of 14 AUs from the CK+ dataset [25] of posed facial expressions, and AU12 and AU14 from the Binghamton dataset [26] of spontaneously displayed facial expressions. A qualitative analysis of AU intensities was reported in the work by Bartlett et al. [27], where distances to the SVM margins, learned for AU detection, were used to obtain (continuous) intensity of AUs.

Overall, from the modeling perspective, the previous work on AU intensity estimation has the following limitations.

1) Modeling the intensity levels on a nominal scale, as in the first group of methods, is suboptimal because models such as the standard SVM [4], [20] treat each intensity level independently, thus, ignoring their total ordering.
2) Modeling the intensity levels on a continuous scale, as in the second group of methods, does not fit the problem well because of the range of each intensity. For example, as shown in Fig. 1, C and D intensity levels cover a larger range of appearance changes than the other levels. Moreover, discrete rating of intensity levels is often preferred and can be accomplished more easily by human coders than the labeling of continuous-valued intensities.
3) The learning/inference is static, i.e, per-frame/window. However, for some AUs, temporal changes in facial expressions carry more discriminative information about the AU intensity than their spatial changes [28], [3].
4) These static methods are not context-sensitive since they answer only the context-question *how* (in terms of changes of facial expressions at the feature level) from the W5+ model. To achieve context-sensitive modeling, i.e., to allow context to influence intensity estimation of AUs, target models need to account for two or more context questions simultaneously. Only then, the models are expected to better disambiguate between the intensity levels of AUs in different contexts.
5) The frequency of occurrence of intensity levels of AUs in spontaneous facial expressions is usually highly skewed toward lower levels (see Fig.4). This data imbalance makes it difficult for the existing models to discriminate accurately between the minority classes (i.e., the higher intensity levels).

The context-sensitive CORF model introduced in this paper addresses the limitations mentioned above. Note also that context modeling has been addressed in other domains such as image annotation (e.g., [29], [30]) or activity recognition (e.g., [31], [32]). These approaches typically model context in terms of co-occurrences of different classes (objects or activities) using CRFs for *nominal* data. By contrast, in our *ordinal* model we employ the more general W5+ context model. To the best of our knowledge, this is the first work that exploits the context in a principled manner, in addition to addressing the other limitations of the existing approaches, in order to improve AU intensity estimation from spontaneously displayed facial expressions.

3 Ordinal Regression

To account for ordinal structure in labels y (i.e., the intensity levels of AUs), different models for ordinal responses can be employed (e.g., see [33]). We adopt the latent variable approach introduced in [14]. In this approach, the ordinal variable y is assumed to be a manifestation of some continuous latent variable z. Then, the noiseless ordinal likelihood is defined as

$$P_{ideal}(y=k|z) = \begin{cases} 1 & \text{if } z \in (\gamma_{k-1}, \gamma_k] \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $k = 1, ..., K$, with K being the number of ordinal responses (in our case, the number of AU intensity levels),

and $\gamma_0 = -\infty \leq \cdots \leq \gamma_K = \infty$ are the thresholds or cut-off points that divide the real line into K contagious intervals. These intervals map the continuous latent variable z to the discrete variable y, which satisfies the monotonicity constraints. The latent variable z is defined as:

$$z = \beta^T x + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

where $x \in \Re^D$ is a D-dimensional covariate vector, β is the ordinal projection vector, and ϵ is a Gaussian noise with zero mean and variance σ^2. Then, the ordinal likelihood is constructed by contaminating the ideal model with noise:

$$\begin{aligned} P(y=k|z) &= \textstyle\int P_{ideal}(y=k|z) \cdot \mathcal{N}(\epsilon; 0, \sigma^2)\ d\epsilon \\ &= \Phi(\lambda_k) - \Phi(\lambda_{k-1}), \end{aligned}, \tag{3}$$

where $\Phi(\lambda) = \int_{-\infty}^{\lambda} \mathcal{N}(\xi; 0, 1) d\xi$ is the normal cumulative distribution function (cdf), and $\lambda_k = \frac{\gamma_k - \beta^T x}{\sigma}$ are the cumulative probits [13]. Usually, σ is set to one for identification purposes [34]. Note that the most critical aspect that differentiates the ordinal regression [13], [14], [35] from multi-class classification [36], [12] is the modeling strategy: while the former learns a single projection (β), which has the same effect on the covariate values of different ordinal responses, the latter learns a separate projection for each response ($\beta_k,\ k = 1, ..., K$). Therefore, the ordinal models are more parsimonious and often more robust, if the responses are indeed of ordinal nature [34].

4 Context-sensitive Conditional Ordinal Random Fields (cs-CORF)

In this Section, we first introduce the concept of context-sensitive modeling of ordinal variables (i.e., the intensity levels of AUs) using the ordinal regression framework described in Sec.3. We then generalize this model by allowing its variance to be a function of the context-sensitive covariates. The resulting model is then integrated into the framework of CRFs to account for temporal dependences between the ordinal variables. We also introduce a weighted softmax-margin learning approach that enables the proposed model to handle skewed distribution of the intensity levels. Lastly, we describe the used regularizers and the inference procedure.

4.1 Latent Variable Approach to Context-sensitive Modeling

The context-sensitive modeling of data is attained by allowing the effects that correspond to different context questions to influence the output responses via the latent variable z. To this end, we define the latent variable model as

$$\begin{aligned} z = \beta_1^T x^{who} + \beta_2^T x^{where} + \beta_3^T x^{what} \\ + \beta_4^T x^{how} + \beta_5^T x^{when} \ + \beta_6^T x^{why} + \epsilon, \end{aligned} \tag{4}$$

where the noise term is defined as in (2). The covariates $(x^{who}, x^{where}, x^{how}, x^{when}, x^{why})$ aim to 'answer' each of the corresponding context questions from the W5+ context model [11]. Note that although z is linear[1] in these covariates, this is not the case with the response variable y as it is *non-linearly* related to z via (3). Therefore, the estimated intensity is the result of non-linear interactions of the context covariates accounting for each context question in the model.

4.2 Modeling of context questions *who* and *how*

To demonstrate how the latent variable model in (4) can be applied to the target task (i.e., AU intensity estimation), in what follows we focus on the context questions *who* and *how*, however, the other context questions can be modeled in a similar manner. These two questions are of particular importance since the first directly accounts for the subject-specific aspect of the context. The second accounts for relationships between the observed facial changes and the corresponding AU intensities, which are assumed to be common to all subjects. To model these two context questions, we introduce the context-related covariate effects (CRE) and the context-free covariate effects (CFE), corresponding to the covariates x^{who} and x^{how} in (4), respectively. The latter are called context-free in this paper as these covariates coincide with those used in the context-free models for the target task. We derive the CRE and CFE components as follows. Given a sequence of ordinal intensities, $\mathbf{y}_i = \{y_{i1}, \ldots, y_{iT_i}\}$, with the corresponding covariate values $\mathbf{x}_i = \{x_{i1}, \ldots, x_{iT_i}\}$, we decompose x_{ij} into CRE ($x_i^u = C^{-1} \sum_{c=1}^{C} x_{ic}$) and CFE ($x_{ij}^r = x_{ij} - x_i^u$) components. The CRE are considered constant across the sequence but may vary between sequences (e.g., the facial shapes of different subjects). Here, we estimate it from the first C neutral intensity frames in a sequence[2]. On the other hand, the CFE account for variability *within* the sequence (i.e., the expression/AU intensity). With these newly introduced effects, we write the latent variable model from (4) as

$$z_{ij} = \beta_u^T x_i^u + \beta_r^T x_{ij}^r + \epsilon_{ij}. \tag{5}$$

By following the same approach as in (3), we obtain the context-sensitive cumulative probits as

$$\lambda_{ijk} = \gamma_k - \beta_u^T x_i^u - \beta_r^T x_{ij}^r,\ k = 1, \ldots K, \tag{6}$$

where $\sigma = 1$. From (6), we can distinguish between (i) an overall effect of the CRE component, as measured by the association of the person-specific biases with the responses, and (ii) the time-varying CFE component within the sequence. Intuitively, the locations of the thresholds γ_k, dividing the ordinal line into the bins corresponding to different intensity levels, are adjusted to the target subject by means of the CRE component ($\beta_u^T x_i^u$). On the other hand, the CFE component ($\beta_r^T x_{ij}^r$) ensures that the intensity-related variation is placed correctly into such adjusted bins. This simultaneous interaction of the CRE and CFE components with the other parameters of the model is at the heart of our approach. If the CRE component is removed from the model ($\beta_u = 0, \beta_r \neq 0$), the context is lost and it may become difficult for the model to adapt to different subjects. On the other hand, assuming the common effects ($\beta_u = \beta_r$) can

1. Non-linear mappings can be obtained as in [37] by applying Representer Theorem to the regularized loss defined in 20. However, in this paper we focus on linear models.

2. We set C=5 to obtain a robust estimate of the target covariates. However, a single frame should suffice.

lead to very misleading association of covariates with the responses, since they model neither CRE nor CFE covariate effects.

Heteroscedastic noise model. The latent variable in (5) is defined using the homoscedastic noise model, i.e., the variance σ^2 of the noise term is constant. However, since the CRE component has an additive effect on the locations of the model's thresholds γ_k within a sequence, it accounts only for the mean level of the subject's expressiveness level. For the model to be able to fully adapt to expressiveness levels of different subjects, we also need to allow the scale of the thresholds to change. This can be attained by relaxing the assumption of constant σ, i.e., by allowing the noise level to vary as a function of covariates. The ordinal models with varying noise levels are usually termed heteroscedastic ordinal models [34]. Thus, we further extend the latent variable model in (5) by introducing separate noise terms

$$z_{ij} = \beta_u^T x_i^u + \beta_r^T x_{ij}^r + \epsilon_i^u + \epsilon_{ij}^r + \epsilon_{ij}, \tag{7}$$

where $\mathcal{N}(\epsilon_i^u; 0, \sigma_u(x_i^u))$ and $\mathcal{N}(\epsilon_{ij}^r; 0, \sigma_r(x_{ij}^r))$. We also keep the constant noise term to account for sources of variation that are not included in the model (e.g., the effects of the other context questions). Because we assume that the three noise terms are independent, the distribution of the overall noise in the model is a zero-mean Gaussian with the variance

$$\sigma^2(x_{ij}) = \sigma_u^2(x_i^u) + \sigma_r^2(x_{ij}^r) + \sigma_o^2. \tag{8}$$

The first two terms on the right represent the CRE and CFE variance, respectively, and are defined as the log-linear function of their covariates, i.e., $\log \sigma_u = \upsilon_u^T x_i^u$ and $\log \sigma_r = \upsilon_r^T x_{ij}^r$. The parameters υ_u and υ_r indicate the importance of the CRE and CFE variances, respectively, and log function ensures that the standard deviation is positive. Using the latent variable model in (8), and after the marginalization in (6), we obtain the context-sensitive cumulative probits, which also have the changing variance, as

$$\lambda_{ijk} = \gamma_k \sigma^{-1}(x_{ij}) - (\beta_u^T x_i^u + \beta_r^T x_{ij}^r)\sigma^{-1}(x_{ij}), \tag{9}$$

where the context-sensitive ordinal likelihood is $P(y_{ij} = k|z_{ij}) = \Phi(\lambda_{ij,k}) - \Phi(\lambda_{ij,k-1})$. From (9), we see that both the constant CRE and time-varying CFE covariates influence the scale of the model's thresholds as well as their location. Note that since we use the same covariates in the location and scale models, the identification may be fragile [34]. Nevertheless, the model can still be identified due to the different functional forms specified for the covariates, but it is necessary to regularize the parameters. This is explained in Sec.4.5.

4.3 Modeling of context question *when*

The context-sensitive ordinal likelihood in Sec.4.2 models the context questions at the feature level by allowing the proposed *covariate* effects to simultaneously influence estimation of the AU intensity via the latent variable z in (4). However, some aspects of these questions can be handled more naturally at the model level. We demonstrate this on the context question *when* by modeling its temporal aspect, i.e., temporal correlation between the AU intensity levels[3]. In this way, we also avoid the potential problem of temporal misalignment of raw image features. Specifically, we employ the modeling strategy of the linear-chain Conditional Random Field (CRF) [12], where the conditional distribution $P(\mathbf{y}_i|\mathbf{x}_i;\theta)$ of a sequence $\{\mathbf{y}_i, \mathbf{x}_i\} = \{(y_{i1}, x_{i1}), \ldots, (y_{iT_i}, x_{iT_i})\}$, $i = 1, \ldots, N$, is represented as the Gibbs form clamped on observations $\mathbf{x}_i$:

$$P(\mathbf{y}_i|\mathbf{x}_i;\theta) = \frac{\exp(\sum_{j=2}^{T_i} \Psi(y_{i,j-1}, y_{ij}, \mathbf{x}_i;\theta))}{\sum_{\bar{\mathbf{y}} \in \mathcal{Y}^{|T_i|}} \exp(\sum_{j=2}^{T_i} \Psi(\bar{y}_{i,j-1}, \bar{y}_{ij}, \mathbf{x}_i;\theta))}, \tag{10}$$

and where T_i is duration of the i-th sequence, and $\mathcal{Y}^{|T_i|}$ is the set of all possible configurations of an output graph $G = (V, E)$. Furthermore, θ are the parameters of the score function $\Psi(y_{i,j-1}, y_{ij}, \mathbf{x}_i;\theta) \equiv \Psi_{ij}(y)$[4] defined on *node* cliques ($r \in V$) and *edge* cliques ($e = (s, r) \in E$) of the graph as

$$\Psi_{ij}(y) = f_n(y_{ij}, \mathbf{x}_i) + f_e(y_{i,j-1}, y_{ij}). \tag{11}$$

The choice of the *node* $f_n(y_{ij}, \mathbf{x}_i)$ and *edge* $f_e(y_{i,j-1}, y_{ij})$ features depends on the target task, and plays a crucial role in the definition of CRFs. We use the introduced context-sensitive ordinal likelihood function to define the *node* features as

$$f_n(y_{ij}, \mathbf{x}_i) = \sum_{k=1}^{K} I(y_{ij} = k) \cdot \log P(y_{ij} = k|z_{ij}), \tag{12}$$

where $P(y_{ij} = k|z_{ij})$ is defined in Sec.4.2, and $I(\cdot)$ is the indicator function that returns 1 (0) if the argument is true (false). On the other hand, the *edge* features model the first order Markov dependence between the ordinal responses as

$$f_e(y_{i,j-1}, y_{ij}) = \sum_{m,k=1}^{K} I(y_{i,j-1} = m \wedge\ y_{ij} = k) \cdot u_{mk}, \tag{13}$$

where u_{mk} measures the temporal association between the ordinal responses. Note that the denominator in (10) guarantees that the probability sums to one, and is computed using (12) and (13), but without the indicator function. Now, given training data pairs $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^N$, parameters $\theta = \{\{\gamma_k\}_{k=1}^{K-1}, \sigma_o, \beta_u, \beta_r, \upsilon_u, \upsilon_r, \{u_{mk}\}_{m,k=1}^K\}$ are found by minimizing the penalized *log-likelihood* function:

$$\min_\theta -\sum\nolimits_{i=1}^{N} \log P(\mathbf{y}_i|\mathbf{x}_i;\theta) + R(\theta), \tag{14}$$

where $R(\theta)$ is the regularization term that prevents the model overfitting. We name the model with the objective function in (14) the context-sensitive Conditional Ordinal Random Field (cs-CORF).

3. However, there are other aspects of this question such as *when* different AUs co-occur, which can be accounted for more efficiently at the feature level using the model in (4). For instance, this can be accomplished by using outputs of S independently trained AU detectors to form $x^{when} = \{\text{AU}_i\}_{i=1}^S$, where $\text{AU}_i = 1$ if AU_i is active, and $\text{AU}_i = 0$ otherwise, and by plugging it into (4). Although accounting for these co-occurrences at the model level is possible (e.g., using factorial CRFs[38]), this would increase considerably the learning and inference complexity of the model.

4. We drop dependence on $j-1$, $\mathbf{x}_i$ and θ for notational simplicity.

Note that temporal models such as CORF [39] and linear-chain CRF for nominal data [12], can also be considered as context-sensitive since they answer the context questions *how* and *when* simultaneously, via their node and edge potentials. However, these models are not fully context-sensitive in the W5+ sense as they do not provide explicit means for modeling the other four context questions (*who*, *where*, *what*, and *why*). On a more subtle level, these temporal models fail to account for heteroscedastic variance in their node potentials. Yet, this is important for capturing non-linear effects of the context covariates on the AU intensity levels. All this is successfully accounted for in the proposed cs-CORF model.

4.4 Weighted Softmax-margin Learning

To deal with skewed distribution of ordinal responses, we relate the large-margin learning approach for sequence classification in [40] to the CRF model in (10). However, in contrast to [40], we introduce scaling of the slack variables, which induces a higher penalty when making errors on minority classes during learning. We start from standard primal learning approach for max-margin models [17], [18]:

$$\begin{array}{l}\min\limits_{\zeta_{ij},\theta} R(\theta) + \sum_{i=1}^{N} \sum_{j=2}^{T_i} \zeta_{ij} \\ s.t.\ \Psi_{ij}(y) - \Psi_{ij}(\bar{y}) \geq \Delta_{ij}(y,\bar{y}) - \frac{\zeta_{ij}}{w_{ij}(y,\bar{y})}, \\ \forall \bar{y} \in \mathcal{Y}, \zeta_{ij} > 0, i = 1 \ldots N, j = 2 \ldots T_i,\end{array} \tag{15}$$

where the large-margin set of constraints are applied to the score function defined in (11). These constraints enforce the difference between the scores of the correctly labeled cliques ($\Psi_{ij}(y)$) and incorrectly labeled cliques ($\Psi_{ij}(\bar{y}), y \neq \bar{y}$) to be greater than the loss $\Delta_{ij}(y,\bar{y})$. This loss is defined on the temporally neighboring pairs of labels as the weighted Hamming loss, i.e., $\Delta_{ij}(y,\bar{y}) = 1 - [\alpha I(y_{ij} = \bar{y}_{ij}) + (1-\alpha) I(y_{i,j-1} = \bar{y}_{i,j-1})]$, for j>1 and $0 \leq \alpha \leq 1$, where for j=1 we set α=1. The weighting of the slack variables ζ_{ij} is done using the weights derived based on the prior distribution of the intensity levels as $w_{ij}(y,\bar{y}) = w_{ij}(y) = 1/(p(y_{ij})+\varepsilon)$, where $p(y_{ij}) = n_{y_{ij}} / \sum_{k=1}^{K} n_k$. Here, n_k is the number of training examples with intensity level $k \in \{1 \ldots K\}$, and ε is chosen from the range $[0,1]$ to avoid minority classes dominating the overall loss. The constraints in (15) can further be written as

$$w_{ij}(y)\Psi_{ij}(y) - w_{ij}(y)(\Psi_{ij}(\bar{y}) + \Delta_{ij}(y,\bar{y})) \geq -\zeta_{ij}. \tag{16}$$

Note that when the weight $w_{ij}(y)$ is set to one, the constraint in (16) is equivalent to that used in the conventional n-Slack large-margin learning with margin-rescaling [17]. We now re-write the optimization problem in (15) in the form that folds the multiple constraints into a single constraint per training sequence as

$$\begin{array}{l}\min\limits_{\zeta_i,\theta} R(\theta) + \sum_{i=1}^{N} \zeta_i \\ s.t.\ \sum_{j=2}^{T_i} \left[\Psi_{ij}^{w}(y) - (\Psi_{ij}^{w}(\bar{y}) + \Delta_{ij}^{w}(y,\bar{y}))\right] \geq -\zeta_i, \\ \forall \bar{\mathbf{y}}_i \in \mathcal{Y}^{|T_i|},\ i = 1 \ldots N,\ \zeta_i > 0,\end{array} \tag{17}$$

where we simplify the notation by defining $\Psi_{ij}^{w}(y) \equiv w_{ij}(y)\Psi_{ij}(y)$, $\Psi_{ij}^{w}(\bar{y}) \equiv w_{ij}(y)\Psi_{ij}(\bar{y})$ and $\Delta_{ij}^{w}(y,\bar{y}) \equiv w_{ij}(y)\Delta_{ij}(y,\bar{y})$. While the optimization problem (OP) in (17) has $N \cdot \mathcal{Y}^{|T_i|}$, $i = 1...N$, constraints, one for each possible combination of labels $\bar{\mathbf{y}}_i = (\bar{y}_{i1}, ..., \bar{y}_{iT_i}) \in \mathcal{Y}^{|T_i|}$, it has only one slack variable ζ_i per sequence. This is exactly what we need for sequence learning since, in contrast to ζ_{ij} in OP in (15), each ζ_i in OP in (17) can now be optimized individually for given θ. The smallest feasible ζ_i given θ is then

$$\zeta_i = \max_{\bar{\mathbf{y}}_i \in \mathcal{Y}^{|T_i|}} \sum_{j=2}^{T_i} (\Psi_{ij}^{w}(\bar{y}) + \Delta_{ij}^{w}(y,\bar{y})) - \sum_{j=2}^{T_i} \Psi_{ij}^{w}(y). \tag{18}$$

We next obtain a more workable constraint by replacing the max term with the $softmax$ upper bound using the inequality $\max_i g_i \leq \log \sum_i e^{g_i}$, which leads to

$$\zeta_i = \log \sum_{\bar{\mathbf{y}}_i \in \mathcal{Y}^{|T_i|}} \mathrm{e}^{\sum_{j=2}^{T_i} \Psi_{ij}^{w}(\bar{y}) + \Delta_{ij}^{w}(y,\bar{y})} - \sum_{j=2}^{T_i} \Psi_{ij}^{w}(y) \tag{19}$$

The constraint in (19) is more restricted than that in (18) since it uses an upper bound on the gap between the scores of the true and model labeling of the sequence. More importantly, in contrast to the max constraint, the $softmax$ large-margin constraint is a differentiable function of the model parameters. We use this to cast the OP in (17) as an unconstrained OP. Specifically, since the constraint in (19) has a form similar to that of the negative log of the conditional probability of CRFs defined in (10), we can formulate the weighted softmax-margin learning of the CRF/cs-CORF model as the following (unconstrained) OP:

$$\min_{\zeta_i,\theta} R(\theta) + \sum_{i=1}^{N} \zeta_i \equiv \min_{\theta} R(\theta) - \sum_{i=1}^{N} \log P^{w}(\mathbf{y}_i|\mathbf{x}_i;\theta), \tag{20}$$

where the conditional likelihood-like term P^w is defined as

$$P^{w}(\mathbf{y}_i|\mathbf{x}_i;\theta) = \frac{\exp(\sum_{j=2}^{T_i} \Psi_{ij}^{w}(y))}{\sum\limits_{\bar{\mathbf{y}} \in \mathcal{Y}^{|T_i|}} \exp(\sum_{j=2}^{T_i} \Psi_{ij}^{w}(\bar{y}) + \Delta_{ij}^{w}(y,\bar{y}))} \tag{21}$$

Note that OP in (20) has a form similar to that of the related softmax-margin approaches (e.g., [40], [41], [17], [18]). However, none of those approaches addresses the problem of class imbalance. Note also that 'slack-rescaling' in [17], [18] is defined as another way, in addition to 'margin-rescaling', of large-margin structured learning, where the slack variables are scaled using the inverse *loss* $\Delta(y,\bar{y})$. This is different from our approach where the slack variables are scaled with the inverse *weights* $w(y)$ in order to balance the contribution of the loss on the minority and majority classes. Moreover, we include the *loss* $\Delta(y,\bar{y})$ using the 'margin-rescaling' approach because, in contrast to 'slack-rescaling', it allows us to formulate the OP as that of standard CRFs (with the likelihood-like term in 21).

4.5 Regulizers

To deal with the order constraints in threshold parameters γ, we introduce the displacement variables η_k, where $\gamma_j = \gamma_1 + \sum_{k=1}^{j-1} \eta_k^2$ for $j = 2, \ldots, K-1$. So, γ is replaced by the unconstrained parameters $\{\gamma_1, \eta_1, \ldots, \eta_{K-2}\}$. Another important issue is the regularization of the parameters of the cs-CORF model. We use the L_2 regularizer for standard CRF

parameters, resulting in the regularization term $R(\theta)$ as:

$$R(\theta) = \rho_1(\|\beta_u\|^2 + \|v_u\|^2) + \rho_2(\|\beta_r\|^2 + \|v_r\|^2) + \rho_3\|u\|^2, \quad (22)$$

where (ρ_1, ρ_2, ρ_3) are the regularization parameters, which help to avoid the model overfitting by controlling the impact of the CRE and CFE effects as well as of the dynamics in the model. The optimal parameters θ are then found by minimizing the objective in (20) with the quasi-Newton LBFGS method. The regularization parameters are found using a cross validation procedure, as explained in Sec.5. The inference of test sequences is performed by Viterbi decoding, applied to the 'unweighted' conditional likelihood in (10).

5 EXPERIMENTS

5.1 Datasets and Experimental Procedure

Datasets. Evaluation of the proposed model is performed on the UNBC-MacMaster Shoulder Pain Expression Archive (Shoulder-Pain) dataset [19] and the Denver Intensity of Spontaneous Facial Actions (DISFA) dataset [20]. To the best of our knowledge, these are the only two sets of naturalistic data that contain a large number of FACS coded AUs and their intensity. We denote these intensity levels using ordinal scores: 0 (not present) to 5 (maximal intensity).

The Shoulder-pain dataset contains video recordings of 25 patients suffering from chronic shoulder pain while performing a range of arm motion tests. A total of 200 image sequences were recorded. The coding of 11 AUs (4, 6, 7, 9, 10, 12, 20, 25, 26, 27 and 43) and their intensity is provided for each frame. As there are only a few examples of higher intensities of AU27, we do not include this AU in our experiments. For similar reasons, we merge examples of levels 4 and 5 of AU12 and of AU20. We also use the intensity coding of pain to evaluate the proposed model. Pain is regarded a high level facial event, and its intensity is defined on a 0-15 ordinal scale using Prkachin and Solomon formulae ($pain = AU4 + \max(AU6, AU7) + \max(AU9, AU10) + AU43$) [42]. As there are only few examples of high intensity of pain (see [19] for details), we grouped the intensity levels as: 0(0), 1(1), 2(2), 3(3), 4-5(4), and 6-15(5).

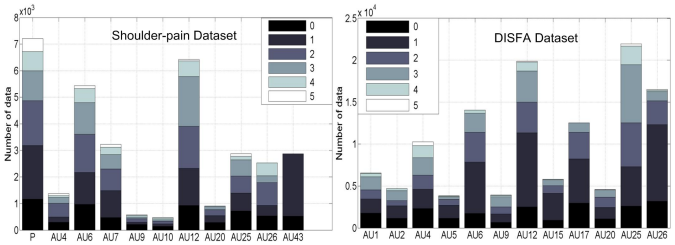


Fig. 4: Distribution of intensity levels in the AU data used from the Shoulder-pain (*left*) and DISFA (*right*) datasets.

The DISFA dataset contains video recordings of 27 subjects watching YouTube videos. Each image frame was coded in terms of 12 AUs (1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25 and 26) and their intensity. Since for AU15 and AU20, there are no examples of the intensity level 5 and only a few examples of level 4, we merged levels 3 and 4, resulting in 4 intensity levels for these AUs. For the same reason, we merged examples of the intensity levels 4 and 5 for AU17. Examples of AUs that are present in either the Shoulder-pain or DISFA dataset, or both, are shown in Fig.3. Since the recordings contain predominantly expressionless faces (i.e., 0 intensity level for all AUs), the sequences from both datasets were pre-segmented per AU. Specifically, the segments containing non-neutral AU intensity were marked first. Then, the surrounding neutral-intensity frames were added at the beginning and end of these segments. The number of 'neutral' frames was balanced with the second most frequent intensity level of the target AU. Fig.4 shows the distribution of the intensity levels after segmentation of the sequences. The sequences made in this way were used to evaluate the models.

Features. As input to our model, we used the facial representation based on geometric features (i.e., the locations of 66 facial landmarks depicted in Fig.8, and obtained using a 2D Active Appearance Model (2D-AAM) [19]). We chose these features as they have already shown good performance in variety of AU recognition tasks (e.g.,[43], [5]). Note, however, that in [43], [21] the authors showed that improved recognition performance can be attained when both geometric and appearance features (e.g., gray-scale intensity) are used. Yet, registration of facial appearance is challenging because of large head movements typically present in spontaneous facial data. While this can partly be addressed as in [44] by engineering pose-robust appearance-based features, here we limit our consideration to the geometric features. To register the features, we applied an affine transform that maps the facial landmarks from faces in each dataset to those of the corresponding reference face (we used the average face from the target datasets). To reduce the number of the features, we applied Principal Component Analysis (PCA) to 132-D feature vectors obtained by concatenation of (x, y) coordinates of the 66 facial landmarks. On average, this resulted in 18-D features, preserving 97% of data variance. These were then used to derive the CRE and CFE covariates, as explained in Sec.4.1.

Models. We compare the performance of the cs-CORF and CORF models, and their variants. Specifically, we compare the maximum-likelihood and the proposed weighted softmax-margin learning of the models, denoted by 'ml' and 'w', respectively. Next, we compare the CORFs with the homoscedastic (σ=1) and heteroscedastic ($\sigma(x)$) noise models, with the latter denoted by 'h'. To compare the ordinal with nominal modeling of the target tasks, we show the performance of standard linear-chain CRF model [12], trained using both 'ml' and 'w' learning. As the baseline model, we use one-vs-all SVM. We also perform comparisons with the state-of-the-art *static* ordinal regression models, Support Vector Ordinal Regression (SVOR) with implicit constraints [35], and Gaussian Process Ordinal Regression (GPOR) with the Laplace approximation [14]. In the kernel methods (SVM/SVOR/GPOR), we used linear kernel function, to have a fair comparison with the linear CRF/CORF-based models. Finally, we include the comparisons with the state-of-the-art

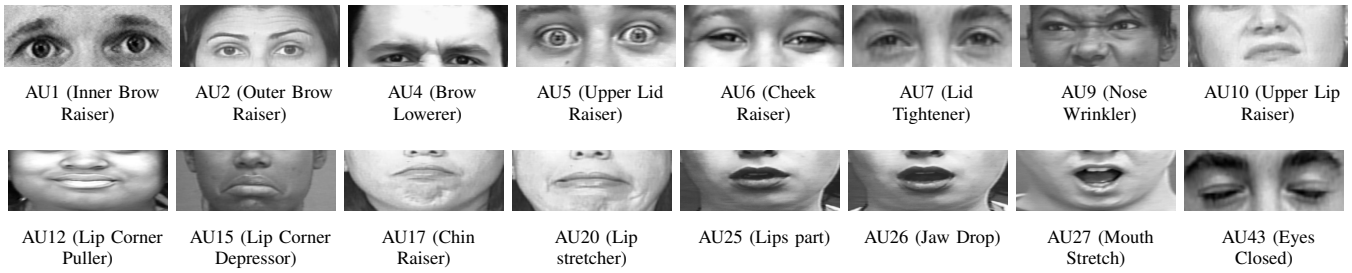


Fig. 3: Examples of AUs available in Shoulder-Pain and DISFA datasets. The images are obtained from http://www.cs.cmu.edu/~face/facs.htm.

models for AU intensity estimation: the RVM approach [21], where continuous estimation of AU intensity is performed, and Spectral Regression [23] combined with one-vs-one SVM (SR+SVM) [4], [20]. The continuous predictions by the RVM-based approach were rounded to the nearest intensity level. For the SR+SVM approach, AU-specific subspaces were selected by running a validation procedure on the training set. In both methods, we used the RBF kernel, as used in the original works [21], [20]. The width of the RBF kernel was set as the median of the (feature's) distance set, i.e., $\{\|x_i - x_j\|, i, j = 1, ..., N, i < j\}$ [33]. The hyper/regularization-parameters of all methods were selected by a 5-fold cross validation on the training set using a grid-search in a range $\rho = \{10^{-4}, 10^{-3}, ..., 1, 2, 5\}$. If not stated otherwise, in all our experiments we applied a 5-fold cross validation procedure, with each fold containing intensity sequences of different subjects.

Evaluation Scores. We use the following scores:

F1. This is the standard score for nominal classification. We report the average score computed as $\text{F1} = \frac{1}{K}\sum_{j=1}^{K}\text{F1}^{(j)}$, where $\text{F1}^{(j)}$ is the score for class $j = 1, \dots K$, and K is the number of classes (i.e., the AU intensity levels).

Mean Absolute Error (MAE). This score is commonly used to measure regression and ordinal classification performance [39], [14]. Because of the imbalanced data, we use the weighted version of this score, defined as $\text{MAE} = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{N_k}\sum_{y_i \in N_k}|y_i - \bar{y}_i|$ where N_k is the number of examples from class k, and y_i and $\bar{y}_i$ are the true and predicted class labels, respectively.

Intra-class Correlation (ICC). This is a measure of correlation or conformity of data with multiple targets. It is commonly used in behavioral research to quantify agreement/consistency between different raters [45]. Depending on how the ratings are obtained, different types of this score should be used (see [45] for details). We use the ICC(3,1) model that is based on a Mixed Model ANOVA, with J judges, treated as fixed effects, and N targets, considered as random effects. In our case, $J = 2$ (the true and predicted values), and N is the total number of test examples. The ICC(3,1) is computed as $ICC = \frac{BMS - EMS}{BMS + (J-1)EMS}$, where $BMS = \frac{BSS}{N-1}$ is between-class mean squares and $EMS = \frac{ESS}{(J-1)(N-1)}$ is residual mean squares. BSS and $ESS = WSS - RSS$ are defined as between target sum squares and residual sum of squares, while WSS and RSS are within-target and between raters sum squares, respectively. This score ranges from 0 to 100 (in %), but sometimes negative values can occur [45].

Ordinal Classification Index (OCI). This score is obtained directly from a confusion matrix (CM). Given a normalized CM, OCI [46] is defined as

$$OCI = \min\left\{1 - \frac{\sum_{(r,c)\in path} n_{r,c}}{100 \cdot K + \sum_{\forall(r,c)} n_{r,c}\,|r-c|} + \beta \sum_{(r,c)\in path} n_{r,c}\,|r-c|\right\}$$

, where $n_{r,c}$ is the fraction (in %) of examples from the r-th class predicted as being from the c-th class, and the $path$ is defined as a sequence of entries where two consecutive entries in the path are 8-adjacent neighbors (see [46] for details). For small values of β (we use 0.25), OCI focuses on measuring ordinal performance from CMs. This score is a *dissimilarity* measure ranging from 0 to 100 (in %).

We use these scores because they capture complementary information about the models' performance. Furthermore, all the scores defined above, except ICC, are robust to class imbalance, which makes them suitable for our data.

5.2 Experimental Results

In this section, we first show some qualitative results. We then show the comparisons with the state-of-the-art models using the context-related and context-free covariates. We continue by showing the results for the intensity estimation of *pain* and individual *AUs* from the two datasets, followed by analysis of the models' performance on two specific AUs (6&25). Lastly, we show the results of the cross-dataset experiments.

Qualitative results. To get an insight into the role of the different effects in the proposed model, we first focus on comparisons between the cs-CORF model (CRE and CFE effects) and the homoscedastic CORF model (CFE effects). Both models were optimized using the introduced weighted softmax-margin approach. The performance of the models is demonstrated on the *pain* intensity estimation task using two example sequences. As can be seen from Fig.5 (top row), the predictions by the cs-CORF model are better aligned with the ground truth than those by the CORF model, which fails to correctly guess level 4 in the first sequence, and level 1 in the second. The middle row of Fig. 5 shows the values of the corresponding ordinal projections, along with the model parameters. By looking at the ordinal thresholds of the two models, we see that their scaling, due to the

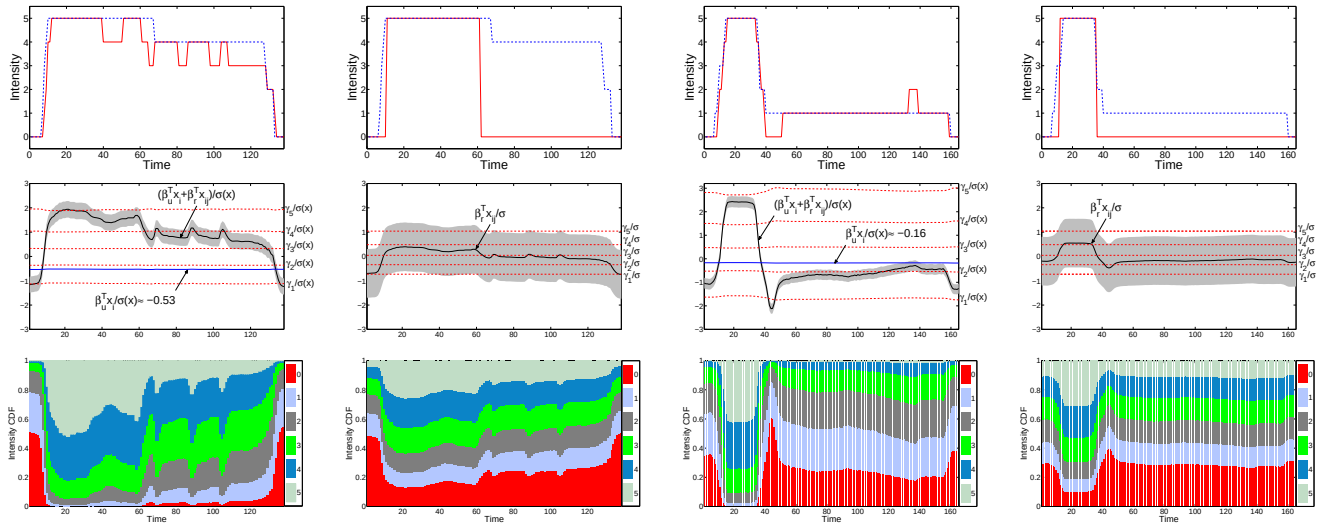


Fig. 5: The intensity estimation of pain from two example sequences of facial expressions from the Shoulder-pain dataset, attained by cs-CORF(w+h) and base CORF(w). The upper row shows true (*dashed blue*) and predicted (*solid red*) labels by the two models. The middle row shows the ordinal projections of the inputs (*solid black*), with their standard deviation σ (*grey*), and the scaled thresholds (*dashed red*). For cs-CORF(w+h), we also plot the context-induced 'bias' (*solid blue*). The bottom row shows the probability of each pain intensity level per frame.

		SVM	GPOR	SVOR	RVM	SR+SVM	CRF(ml)	CRF(w)	CORF(ml)	CORF(w)	CORF(ml+h)	CORF(w+h)
F1	CRE+CFE	24.1 (18.1)	24.4 (17.6)	26.1 (14.2)	24.5 (17.2)	25.7 (15.8)	29.3 (11.9)	32.0 (8.1)	33.2 (6.8)	35.5 (3.9)	35.3 (3.9)	**38.7 (1.4)**
	CFE	24.8 (15.7)	23.5 (20.4)	25.2 (15.1)	27.3 (14.8)	29.7 (11.0)	29.5 (12.1)	31.5 (9.1)	31.0 (10.4)	33.2 (7.2)	32.8 (7.7)	34.8 (4.9)
MAE	CRE+CFE	1.13 (20.3)	0.96 (16.6)	0.88 (13.3)	0.94 (15.5)	0.94 (15.4)	0.87 (12.1)	0.84 (10.3)	0.77 (5.9)	0.73 (3.6)	0.74 (3.9)	**0.69 (1.9)**
	CFE	1.02 (18.5)	1.06 (19.5)	0.91 (14.9)	0.93 (15.7)	0.82 (9.1)	0.86 (11.9)	0.83 (10.3)	0.79 (8.1)	0.78 (7.4)	0.78 (7.5)	0.76 (5.8)
ICC	CRE+CFE	34.9 (17.9)	38.6 (14.9)	41.8 (13.8)	24.7 (20.4)	27.6 (18.7)	45.3 (11.7)	49.7 (7.9)	52.6 (5.9)	54.8 (3.9)	56.0 (2.9)	**59.1 (1.1)**
	CFE	36.9 (16.2)	37.2 (15.5)	38.5 (15.2)	31.5 (17.9)	38.7 (15.0)	46.8 (11.2)	50.2 (7.7)	48.4 (9.9)	50.2 (8.6)	51.2 (7.6)	53.3 (5.5)

TABLE 1: Average performance of the models tested on 23 intensity estimation problems (*pain* + 10 AUs from Shoulder-pain dataset and 12 AUs from DISFA dataset). The numbers in brackets are the average ranks of the models, where the ranking is performed on 46 (=23×2) tasks, as each model is tested using two sets of covariates: the context (CRE+CFE) and context-free (CFE) covariates. The models are ranked for each task separately, the best performing model getting the rank of 1, the second best rank 2, etc. Note that for all three scores, the top ranked model is the proposed context-sensitive CORF(w+h) model (i.e., CORF(w+h) with CRE+CFE).

modeling of the heteroscedastic noise in cs-CORF, results in a better estimation of the intensity levels. However, this scaling cannot fully account for the subject-specific biases. For this, the context (subject) induced bias (CRE) acts in concert with the scaling of the thresholds. Consequently, the partitioning of the target signal into discrete intensity levels is context (subject) dependent. On the other hand, the base CORF model is far less flexible due to its limited parametrization ($\sigma = 1$ and there is no modeling of the context), resulting in poor estimation of intermediate intensity levels. Fig.5 (bottom row) shows that the probability of each intensity level, computed using Eq.(3), is consistent with the models' predictions. From these probabilities, we also conclude that cs-CORF is more discriminative than base CORF.

Comparisons with the state-of-the-art models. Table 1 shows the average results of various models, obtained using 5-fold cross-validation, for 23 intensity estimation tasks including *pain*, 10 AUs from the Shoulder-pain dataset, and 12 AUs from the DISFA dataset. The models were evaluated using two sets of covariates: context (CRE+CFE) and context-free (CFE). In the case of CRE+CFE, the resulting 36D feature vector (obtained by concatenation of the CRE and CFE covariates) was used as input to tested models. To ensure that the performance of the models is consistent on all 46 tasks (i.e., 23 tasks×2 sets of covariates), we performed ranking of the models as in [47], cf. Sec.3.2.2. Specifically, the models were first ranked per task, the best performing model getting the rank of 1, the second best rank 2, etc. In the case of ties, average ranks were assigned. The final ranking was then obtained by averaging the ranks over all tasks. From Table 1, we observe that the base SVM model is outperformed by the SR+SVM model when the context-free covariates are used. This is attributed in part to the fact that the latter performs non-linear feature selection by means of SR, and in part to the fact that it uses a non-linear kernel function in the SVM classifier, as well as one-vs-one learning strategy. This is in contrast to the base SVM model that employs a liner kernel and one-vs-all strategy. On the other hand, both models underperform when the context covariates are used, possibly due to the overfitting of the CRE covariates. The RVM method, although designed for continuous estimation, shows the performance (in terms of F1 and MAE) comparable to that of SVM. However, its ICC scores are lower, which indicates that its estimation of the intensity levels is not always consistent. The static ordinal

		The Shoulder-pain dataset											The DISFA dataset										
		P	AU4	AU6	AU7	AU9	AU10	AU12	AU20	AU25	AU26	AU43	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26
F1	cs-CORF(w+h)	**41.0**	35.0	**41.0**	**38.0**	**45.0**	50.0	**39.0**	36.0	34.0	30.0	**89.0**	**39.0**	41.0	**37.0**	**37.0**	**36.0**	36.0	**38.0**	**37.0**	**44.0**	41.0	**45.0**	**32.0**
	CORF(w+h)	35.0	32.0	36.0	30.0	41.0	49.0	35.0	34.0	33.0	27.0	78.0	35.0	38.0	34.0	33.0	31.0	33.0	34.0	33.0	40.0	37.0	39.0	27.0
	CRF(w)	30.0	27.0	29.0	29.0	33.0	42.0	32.0	32.0	29.0	26.0	76.0	30.0	34.0	30.0	33.0	28.0	29.0	34.0	33.0	37.0	35.0	36.0	25.0
	RVM	22.8	26.7	22.2	22.1	23.5	43.0	27.8	25.5	22.1	22.0	70.7	29.6	29.6	30.7	26.1	27.3	23.3	32.6	24.8	29.7	28.6	34.9	25.9
	SR+SVM	29.4	24.3	23.9	22.3	32.6	43.4	26.7	29.6	36.0	32.4	78.3	30.7	27.0	28.0	27.1	25.3	30.6	26.5	29.0	29.3	34.3	40.4	24.9
MAE	cs-CORF(w+h)	**0.82**	**0.79**	**0.71**	**0.76**	0.70	0.36	**0.68**	**0.74**	**0.81**	1.19	**0.05**	0.80	0.70	**0.82**	0.58	0.60	0.78	0.61	0.50	**0.51**	0.72	**0.57**	0.53
	CORF(w+h)	0.93	0.88	0.79	0.90	0.75	0.41	0.81	0.83	0.95	1.23	0.11	0.85	0.75	0.90	0.63	0.63	0.78	0.60	0.54	0.60	0.83	0.64	0.52
	CRF(w)	1.16	0.99	0.98	1.00	0.82	0.53	0.94	0.93	0.99	1.23	0.13	0.92	0.95	1.02	0.62	0.70	0.91	0.57	0.50	0.60	0.74	0.68	0.56
	RVM	1.00	1.05	1.16	1.25	1.30	0.64	0.98	0.99	1.16	1.50	0.18	0.94	0.82	1.07	0.77	0.72	1.02	0.63	0.64	0.72	0.84	0.68	0.70
	SR+SVM	1.00	0.93	0.97	1.13	0.85	0.63	0.81	0.85	0.97	1.39	0.11	0.88	0.77	1.07	0.60	0.65	0.74	0.69	0.52	0.59	0.77	0.63	0.51
ICC	cs-CORF(w+h)	**64.0**	75.0	**67.0**	**68.0**	63.0	66.0	**62.0**	**47.0**	**58.0**	**38.0**	**73.0**	**61.0**	**68.0**	**67.0**	51.0	**57.0**	58.0	66.0	51.0	**46.0**	49.0	**78.0**	**40.0**
	CORF(w+h)	59.0	72.0	60.0	59.0	61.0	65.0	57.0	39.0	50.0	25.0	61.0	56.0	63.0	63.0	47.0	49.0	55.0	63.0	49.0	38.0	41.0	72.0	30.0
	CRF(w)	58.0	66.0	52.0	54.0	52.0	49.0	51.0	37.0	43.0	29.0	54.0	52.0	55.0	60.0	49.0	48.0	53.0	65.0	44.0	38.0	50.0	72.0	28.0
	RVM	43.1	33.9	18.8	28.9	-0.5	39.1	27.7	16.3	21.7	16.8	46.0	33.9	53.7	44.7	9.8	33.1	35.1	57.3	25.1	26.7	30.9	66.4	31.0
	SR+SVM	44.4	54.6	36.0	27.2	43.4	37.8	34.0	35.2	38.8	18.2	59.1	53.2	46.8	51.9	26.5	26.1	52.2	40.2	20.7	25.5	47.7	69.0	21.4

TABLE 2: The performance of the models on intensity estimation of *pain* (P) and 11 AUs from the Sholder-Pain dataset, and 12 AUs from the DISFA dataset. The results are the averages of the 5-fold cross-validation procedure. We use bold face to indicate that the proposed cs-CORF(w+h) performs significantly better than the rest of the models, based on the paired t-test with $p = 0.05$.

models, GPOR and SVOR, showed a small improvement in their performance when the context covariates are used. Furthermore, SVOR performed better than the base SVM model across all three scores. The improvement in ICC scores of GPOR and SVOR over nominal static models and RVM, in contrast to the the other two scores, implies that there is a bias in the estimated intensity levels by these ordinal models. Also, the lower performance of GPOR in terms of F1 and MAE is ascribed to its learning being less robust to imbalanced data than that of the max-margin models (i.e., SVOR and SVM). Next, the standard CRF(ml) model performed marginally better than the base SVM in terms of F1. However, its MAE and ICC are much better mainly because of the temporal smoothing of the predicted intensity. On the other hand, the proposed weighted softmax-margin learning improved the performance of the CRF compared to that with 'ml' learning. Yet, there is not much difference when using the context or context-free covariates. However, inclusion of the context covariates in the CORF(ml) model results in an improvement in all three scores. CORF(ml) also outperformed the static ordinal models, GPOR and SVOR, which, evidently, remained affected by temporal variability of the data during learning/inference. Then again, the weighted softmax-margin learning (CORF(w)) and the heteroscedastic noise model (CORF(ml+h)) further enhanced the performance of CORF(ml). Moreover, based on the three scores and the ranking of the models, the combination of the weighted learning and the heteroscedastic noise model in cs-CORF(w+h) (i.e., CORF(w+h) with CFE+CRE) is, evidently, the most effective for the target tasks.

Performance on intensity estimation of pain and AUs. Table 2 shows results of the cs-CORF(w+h), CORF(w+h) and CRF(w) models. We also include the results obtained by two state-of-the-art (context-free) models for AU intensity estimation: SR+SVM [20] and RVM [21]. The numbers with bold face in the table indicate that the differences in scores by the proposed cs-CORF(w+h) and the rest of the models are significant, based on the paired t-test ($p = 0.05$). The proposed cs-CORf(w+h) model performs similarly or better than rest of the models on most the tasks. Specifically, from Table 2, in the case of AU12, cs-CORF(w+h) consistently outperforms the other models. We ascribe this to the fact that AU12 involves activation of an oblique muscle, which is characterized by curved motion that is usually subject-specific. Therefore, modeling the context through subject adaptation, obviously results in a better performance than that attained by the context-free models. By contrast, AU10 involves activation of vertically set muscles above the upper lip. Similarly, AU9 involves a vertical pull of the muscles around the nose, which wrinkles the nose and pulls the nostril wings straight up. Due to the subtlety of these facial movements in naturalistic data and the involvement of vertically set muscles (rather than oblique ones), no strong personal characterization is expected in these AUs. Hence, modeling the context does not much improve the intensity estimation of AU9 and AU10. On the other hand, although AU20 involves horizontal motion (elongating the mouth), it often occurs in combination with other AUs (e.g., 10+20+25 or 20+26). Since these combinations are additive, cs-CORF(w+h) separates the facial deformation due to the AU intensity changes, and due to the co-occurring AUs, by means of the CRE effects, resulting in it achieving the better performance on this AU.

Note also that the activation of AU6 wrinkles the skin around the outer corners of the eyes and raises the cheeks. When the facial landmarks are used as features, it becomes impossible to perform detection/intensity estimation of this AU in isolation from other AUs. However, because of the co-occurring AUs, it is still possible to estimate intensity of AU6 (e.g., AU6+AU12, representing genuine smile, frequently co-occurred in the dataset used). Although we do not explicitly model co-occurrences of different AUs, they are implicitly accounted for by the CRE and CFE components. It is also interesting to note that in the case of AU43, the intensity estimation is still better attained by cs-CORF(w+h) than CRF(w), even though this task is binary classification as there only two levels (eyes open/closed). We attribute this to modeling of the context and noise heteroscedasticity in the cs-CORF(w+h). Similarly, in the case of the DISFA dataset (Table 2), the proposed cs-CORF(w+h) achieves the results that are similar or better than those of the other models in most cases. Nevertheless, compared to the Shoulder-pain dataset, some of the differences (e.g., AU12 and AU20) are not significant when $p = 0.05$ is used in the t-test. On the other hand, the

intensity estimation of AU4 (brow lowerer) is much improved. We attribute this to the fact that there are far fewer examples of higher intensity levels of AU4 in the Shoulder-pain than in DISFA dataset, mainly because of the difference in the context stimulus (pain vs. 'YouTube' videos).

Analysis of the intensity estimation performance on AU6 and AU25. To further investigate performance of the models, we choose these two AUs as examples. Note that intensity estimation of AU6 is particularly challenging because it cannot be detected from facial landmarks alone but its inference relies on the feature variation due to the co-occurring AUs. On the other hand, AU25 can be detected from facial landmarks alone (i.e., even when all other AUs are inactive) and is one of the most common facial actions that occurs involuntary in spontaneous facial expressions. Fig.6 shows confusion matrices (CMs) for different models. Also, from each CM, we computed the OCI score, the low values of which indicate good performance (see Sec.5.1). In both cases, the cs-CORF(w+h) estimated the highest intensity levels more accurately compared to the CORF(w+h). By inspecting the CMs of the models, we note that in both the ordinal models most confusion occurred between the neighboring intensity levels. This is in contrast to the rest of the models, which exhibit a more 'dispersed' confusion of the intensity levels, mainly due to the lack of the ordering constraints.

Note, however, that in some cases the ordinal models also confused higher intensity levels with the neutral level. This usually occurs when input features are corrupted by errors in facial landmark localization and/or their registration. To remedy this, one would have to include a mechanism for detecting the source of the problem, i.e., tracking/registration errors. Another reason for confusion of the intensity levels is the large difference in facial morphology of some test subjects and training subjects. The facial features of such test subjects are treated as outliers by the models. Consequently, they easily confuse the intensity levels, sometimes also classifying the whole sequence as having only the neutral intensity levels. The cs-CORF model can deal better with this due to the modeling of the context question *who* as well as heteroscedasticity in the features. However, as we can see from Fig.6, sometimes it also confuses higher intensity levels with the neutral. Again, this occurs when difference between training and test subjects is large. It is also important to mention that in the case of the DISFA dataset, there is a small number of training examples of the highest intensity level of AU25 ($<$30). From Fig.6, the bottom row, we see that none of the models, including cs-CORF, could generalize successfully to this intensity level from the data used.

Fig.7 shows examples of the intensity estimation at the sequence level for the same AUs (6&25). The scores shown in the title of each graph are computed from the depicted sequences. We note that the RVM model estimates correctly the slope but not the scale of the true intensity, which is a consequence of assuming an equal interval scale. On the other hand, because of the classification bias toward the majority classes in the learned subspace, SR+SVM underestimates the intensity levels in most cases. These models are outperformed by the temporal models, with CRF(w) achieving higher F1 compared to that of CORF(w+h). By contrast, CORF(w+h) achieves better MAE and ICC, which are preferable indicators of the intensity estimation performance. However, cs-CORF(w+h) still outperforms these models.

Cross-dataset evaluation of the models. To test the robustness of the models, we perform the cross-dataset evaluation. The models were trained on data from one dataset and tested on data from the other dataset. This is challenging mainly due to: (i) the difficulty of aligning the features between the datasets, (ii) the bias in annotations by different annotators of two datasets, and (iii) the difference in the context stimulus (the pain inducing exercises vs. YouTube videos), which affects the frequency and co-occurrence of AUs, and thus the features to be selected. For this experiment, we used examples of 7 AUs (i.e., 4, 6, 9, 12, 20, 25 and 26) that are present in both datasets. Registration of the facial landmarks between datasets was performed as explained in Fig.8.

From Table 3, we see that the performance of all models is lower for most of AUs compared to that attained on the datasets used to train the models (see Table 2). This is expected because of the reasons (i)-(iii) mentioned above. From Fig.8 we also see that there is a different level of variation in the registered training/test points from the two datasets. This, in turn, negatively affects the models' performance. Furthermore, we note that cs-CORF(w+h) performs similarly to the other models in the case of AU6. Since the context stimulus in the two datasets is quite different, so are the AU co-occurrences, which is important when inferring the intensity of AU6 from the facial points. Therefore, such behavior of the cs-CORF in this particular task is not surprising because this model accounts for the context question *who* (i.e., the subject), and not *why* (i.e., the context stimulus). Also, as we saw before, the estimation of AU20 was not improved significantly with the context modeling, so here it is similar. In the case of AU9 (nose wrinkler), modeling the context helps when training is conducted on DISFA and testing on the Shoulder-pain dataset, but not the other way round. This is caused by inaccuracies in the registration of the facial points around the nose in the latter case (see Fig. 8 on the left), which adversely affected the generalization performance of the model. Nevertheless, in the case of AUs 4, 12, 25 and 26, cs-CORF(w+h) consistently outperforms the other temporal, and static, context-free models. This is also reflected in the average results.

6 Discussion

The results obtained indicate the benefits of the introduced effects in the cs-CORF model for AU intensity estimation. Specifically, the ordinal probit function accounts for the spatial structure in the data, while their temporal structure is accounted for by the edge features. However, when the homoscedastic probit function is used, the model is unable to fully adapt to the varying expressiveness of different subjects. Thus, introducing the context and heteroscedastic effects in the probit model is critical for the model's performance. As evidenced by the results, answering the context question *who* by the inclusion of the CRE component substantially raises the performance of traditional CORF across all three scoring

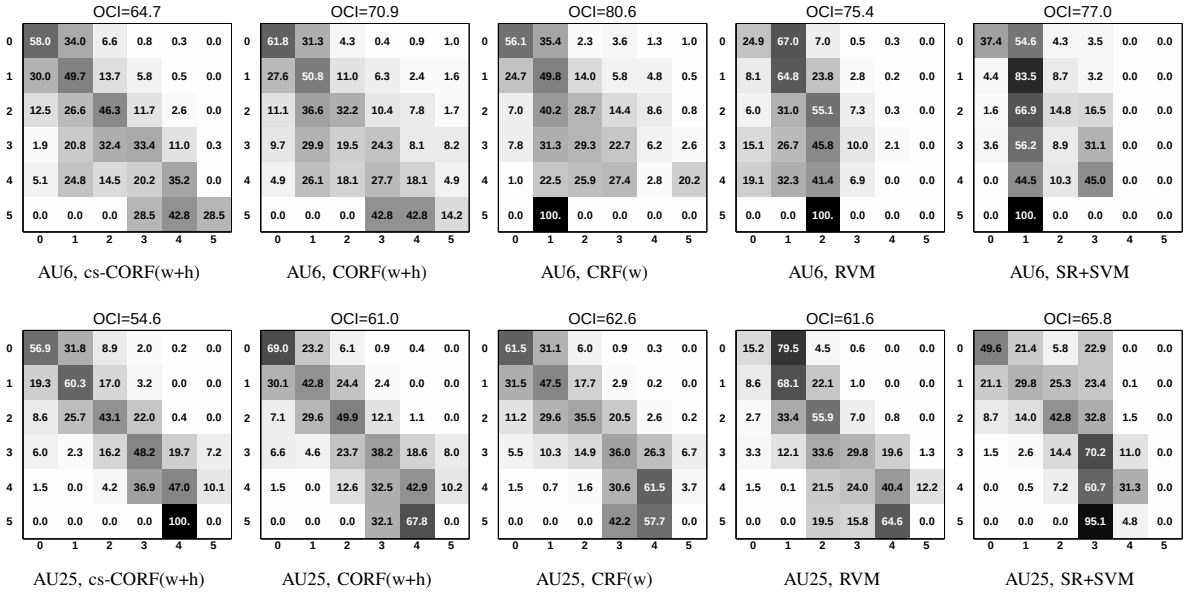


Fig. 6: The (normalized) confusion matrices (CMs) computed from the true and predicted intensity labels, the latter being obtained by the denoted models, for AU6 and AU25 from the DISFA dataset. Note that the lower the OCI score, the better performance.

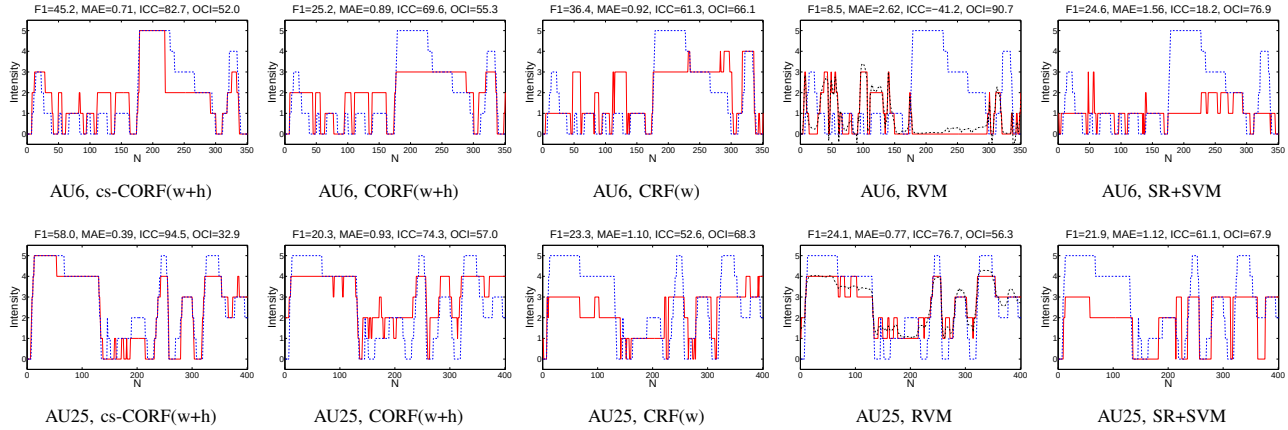
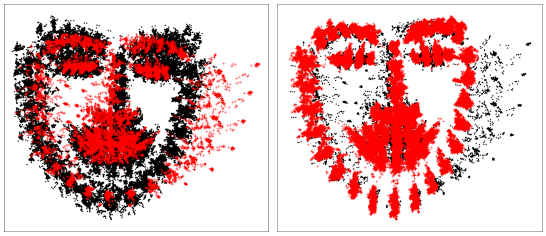


Fig. 7: The true (*dashed blue*) and predicted (*solid red*) intensity of AU6 and AU25 from the DISFA dataset. The sequences shown are obtained by concatenation of several exemplary sequences corresponding to different test subjects. The scores shown at the top of each figure are computed from the depicted sequences. For RVM, we also include the continuous estimation of AU intensity (*dashed black*).



(a) (b)

Fig. 8: Cross-dataset registration: (a) DISFA to Shoulder-pain, and (b) Shoulder-pain to DISFA. The reference face is calculated as the average of the points registered within the datasets (*red*) that are used to train the models. The registered points of the test dataset (*black*) are obtained by using an affine transform that maps the test points to the reference face of the training set.

measures. This is because the CFE component alone is unable to account for the presence of the context but also cannot result in its full removal. This is also true because of the heteroscedastic nature of the data, encoded both in variance and the offset. On the other hand, we conclude that inclusion of the CRE covariates in the non-ordinal models does not improve their overall performance. The main reason for this lies in the lack of parameter tying, i.e., the influence of the CRE and CFE component on each intensity level is modeled independently. By contrast, the CRE- and CFE-related parameters, and the ordinal thresholds in cs-CORF(w+h) act in concert, with the CRE and CFE helping to adjust the location and scale of the thresholds, depending on the input. This allows the model to adapt to the varying expressivity of the subjects, which is reflected by distinct motion patterns (usually oblique) of AUs.

Cross-dataset evaluation			AU4	AU6	AU9	AU12	AU20	AU25	AU26	Av.
F1	cs-CORF(w+h)	SP-D	**28.0**	29.0	25.0	**36.0**	27.0	**37.0**	**19.0**	**28.7**
	CORF(w+h)		25.0	**30.0**	**28.0**	31.0	**31.0**	35.0	14.0	27.7
	CRF(w)		22.0	22.0	**28.0**	34.0	29.0	28.0	16.0	25.5
	RVM		20.0	20.0	21.0	28.0	21.0	33.0	13.0	22.3
	SR+SVM		27.0	19.0	24.0	19.0	16.0	30.0	14.0	21.3
	cs-CORF(w+h)	D-SP	**26.0**	24.0	**39.0**	27.0	38.0	**43.0**	**29.0**	**32.3**
	CORF(w+h)		24.0	24.0	36.0	26.0	**41.0**	38.0	21.0	30.0
	CRF(w)		23.0	22.0	30.0	27.0	33.0	29.0	16.0	25.7
	RVM		22.0	17.0	0.09	24.0	17.0	19.0	16.0	16.4
	SR+SVM		21.0	**26.0**	33.0	**29.0**	31.0	30.0	20.0	27.1
MAE	cs-CORF(w+h)	SP-D	**1.24**	1.25	**1.14**	**0.72**	**0.92**	**0.80**	**1.34**	**1.05**
	CORF(w+h)		1.41	1.24	1.40	0.79	1.08	0.86	1.39	1.17
	CRF(w)		1.59	**1.21**	1.30	0.97	1.06	0.84	1.47	1.21
	RVM		1.57	1.44	1.47	1.05	1.07	0.81	1.62	1.29
	SR+SVM		1.53	1.78	1.54	1.12	1.36	1.13	1.38	1.41
	cs-CORF(w+h)	D-SP	**1.11**	**1.16**	**0.77**	**1.18**	1.04	**0.75**	**1.16**	**1.02**
	CORF(w+h)		1.26	1.25	0.87	1.33	0.95	0.82	1.40	1.13
	CRF(w)		1.20	1.44	1.06	1.34	**1.03**	1.02	1.40	1.21
	RVM		1.24	2.11	2.50	1.38	1.20	1.73	1.42	1.65
	SR+SVM		1.41	1.31	0.99	1.29	1.03	1.31	1.39	1.25
ICC	cs-CORF(w+h)	SP-D	**52.0**	47.0	49.0	**66.0**	**46.0**	**69.0**	27.0	**50.9**
	CORF(w+h)		48.0	**48.0**	**53.0**	62.0	38.0	65.0	28.0	48.8
	CRF(w)		37.0	37.0	44.0	58.0	40.0	57.0	28.0	43.0
	RVM		32.0	34.0	27.0	56.0	25.0	51.0	22.0	35.3
	SR+SVM		41.0	13.0	34.0	44.0	12.0	44.0	**30.0**	31.1
	cs-CORF(w+h)	D-SP	**42.0**	**45.0**	**74.0**	**55.0**	36.0	**62.0**	**27.0**	**48.7**
	CORF(w+h)		37.0	41.0	68.0	50.0	**37.0**	55.0	17.0	43.6
	CRF(w)		37.0	37.0	62.0	41.0	35.0	51.0	15.0	39.7
	RVM		37.0	0.07	0.00	39.0	15.0	25.0	-0.03	16.6
	SR+SVM		25.0	33.0	60.0	39.0	34.0	37.0	26.0	36.3

TABLE 3: Cross-datasets evaluation of the models on 7 AUs present in both datasets. The models are trained using data of target AUs from the Shoulder-pain dataset, and tested on data from the DISFA dataset (denoted as SP-D), and the other way round (denoted as D-SP).

Also, in situations where the facial landmark registration is not well attained and/or a small amount of training data is available (as in the Shoulder-pain dataset), the inclusion of the CRE component increases the robustness of the CORF models. On the other hand, while the CRF nominal model performs rather well (with the inclusion of CRE and CFE covariates), it fails to reach the full performance level of cs-CORF. This is in part due to the lack of ordering constraints on the intensity levels and due to the increased parameter dimensionality. Also, in the ordinal models, the misclassification away from the true 'level' incurs higher cost compared to the level-distance agnostic classification setting (i.e., nominal models). This all leads to more accurate predictions by the proposed cs-CORF. Similar reasoning can be applied to analysis of the performance of other nominal models in the static setting, such as multi-class SVM. Likewise, we showed that the regression model such as RVM is less fit for modeling ordinal data as it assumes the same variability in covariates of different ordinal levels [48].

Also, the traditional methods for sequence classification and AU intensity estimation are designed for balanced data. Yet, because of the imbalanced nature of our data, proper scaling during training is necessary. The most frequent low intensity levels that would otherwise dominate performance scores are properly balanced using the proposed weighted softmax-margin learning for CRFs. This is reflected in improvements of the weighted models (w) over their unweighted counterparts (ml). Lastly, while the standard ordinal models such as GPOR and SVOR provide a solid framework for modeling ordinal data, the class imbalance and the lack of temporal constraints adversely affect their learning and inference. Consequently, they cannot take the full advantage of the context information encoded by the CRE component. This all is successfully accounted for in the proposed cs-CORF model.

7 Conclusions and Future Work

We have proposed a novel approach for context-sensitive modeling of the facial AU intensity levels from spontaneously displayed facial expressions. We addressed limitations of existing approaches that do not leverage the ordinal constraints, and also fail to account for the influence of context on the AU intensity estimation, as well as to account for heterogeneous and imbalanced nature of the data. We showed in our experiments that by accounting for these effects, the proposed context-sensitive model achieves substantially better intensity estimation of AUs and facial expressions of pain.

While in this work we have focused on modeling of the context questions *who*, *how* and *when*, in future it would be interesting to investigate the influence of the other context questions (*where*, *why* and *what*) on the intensity estimation of AUs. There are various ways in which these can be explored within our approach. For instance, the context question *where* can be 'answered' by encoding the subject's head pose via covariates $x^{where} = [\varpi_a \, \varpi_b \, \varpi_c]$, representing pan, tilt and roll angles of the head rotation. These can be obtained using existing methods for pose estimation (e.g., [49]). Likewise, x^{what} can be derived by determining the subject's current focus of attention by means of gaze tracking [50]. Lastly, the context question *why* can be modeled by encoding the subject's emotional states as $x^{why} = \{\text{happy=1, sad=2}, \dots\}$, which can be obtained from target images by applying existing (context-free) classifiers, like in [44]. Studying these and other aspects of target context questions would help to better understand the influence of different contexts on the intensity of facial expressions, and, thus, improve its automated estimation.

Acknowledgments

This work has been supported by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA), and the National Science Foundation under Grant no. IIS0916812.

References

[1] J. Cohn and P. Ekman, "Measuring facial action by manual coding, facial emg, and automatic facial image analysis," in *Handbook of nonverbal behavior research methods in the affective sciences*, 2003.

[2] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Philosophical Transactions of Royal Society B*, vol. 364, pp. 3505–3513, 2009.

[3] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System (FACS): Manual*. A Human Face, 2002.

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

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

[6] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," *IEEE CVPR*, 2013.

[7] Y. Zhu, F. De la Torre, J. F. Cohn, and Y.-J. Zhang, "Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection," *IEEE Trans. on Affective Comp.*, vol. 2, pp. 79–91, 2011.

[8] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE TPAMI*, vol. 29, no. 10, pp. 1683–1699, 2007.

[9] S. Gunnery, J. Hall, and M. Ruben, "The deliberate duchenne smile: Individual differences in expressive control," *J. Nonverbal Behavior*, vol. 37, no. 1, pp. 29–41, 2013.

[10] 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*, vol. 11, no. 5, pp. 310–313, 1998.

[11] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human computing and machine understanding of human behavior: A survey," *Lecture Notes in Artificial Intell.*

[12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *ICML*, pp. 282–289, 2001.

[13] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Stat. Society. Series B*, vol. 42, pp. 109–142, 1980.

[14] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *JMLR*, vol. 6, pp. 1019–1041, 2005.

[15] J. Reilly, J. Ghent, and J. McDonald, "Investigating the dynamics of facial expression," *Lecture Notes in Computer Science*, vol. 4292, pp. 334–343, 2006.

[16] A. Savrana, B. Sankur, and M. Bilgeb, "Regression-based intensity estimation of facial action units," *Image and Vision Computing*, 2012.

[17] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *JMLR*, vol. 6, pp. 1453–1484, 2005.

[18] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural svms," *Mach. Learn.*, vol. 77, no. 1, pp. 27–59, 2009.

[19] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," *IEEE FG*, pp. 57–64, 2011.

[20] S. Mavadati, M. Mahoor, K. Bartlett, P. Trinh, and J. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Trans. on Affective Comp.*, vol. 4, no. 2, pp. 151–160, 2013.

[21] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," *ISVC*, vol. 7432, pp. 368–377, 2012.

[22] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. D. L. Torre, "Continuous au intensity estimation using localized, sparse facial feature space," *IEEE FG*, pp. 1–7, 2013.

[23] D. Cai, X. He, and J. Han, "Spectral regression for efficient regularized subspace learning," *IEEE ICCV*, pp. 1–8, 2007.

[24] A. Savran, B. Sankur, and M. Taha Bilge, "Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units," *Pattern Recogn.*, vol. 45, no. 2, pp. 767–782, 2012.

[25] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," *IEEE CVPR'W*, pp. 94–101, 2010.

[26] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," *IEEE FG*, pp. 1–6, 2013.

[27] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," *IEEE FG*, pp. 223 –230, 2006.

[28] K. L. Schmidt, Z. Ambadar, J. F. Cohn, and L. I. Reed, "Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling," *J. Nonverbal Behavior*, vol. 30, no. 1, pp. 37–52, 2006.

[29] W. Jiang, S.-F. Chang, and A. C. Loui, "Context-based concept fusion with boosted conditional random fields," *IEEE ICASSP*, 2007.

[30] Y. Xiang, X. Zhou, Z. Liu, T.-S. Chua, and C.-W. Ngo, "Semantic context modeling with maximal margin conditional random fields for automatic image annotation," *IEEE CVPR*, pp. 3368–3375, 2010.

[31] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," *IEEE ICCV*, vol. 2, pp. 1808–1815, 2005.

[32] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE TPAMI*, vol. 34, no. 9, pp. 1691–1703, 2012.

[33] T. Kanamori, "Statistical models and learning algorithms for ordinal regression problems," *Journ. of Inf. Fusion*, vol. 14, no. 2, pp. 199–207, 2013.

[34] R. Winkelmann and S. Boes, *Analysis of microdata*. Springer, 2006.

[35] W. Chu and S. S. Keerthi, "New approaches to support vector ordinal regression," *ICML*, pp. 145–152, 2005.

[36] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *JMLR*, vol. 2, pp. 265–292, 2001.

[37] O. Rudovic, V. Pavlovic, and M. Pantic, "Kernel conditional ordinal random fields for temporal segmentation of facial action units," *IEEE ECCV'W*, 2012.

[38] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," *JMLR*, vol. 8, pp. 693–723, 2007.

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

[40] F. Sha and L. K. Saul, "Large margin hidden markov models for automatic speech recognition," *NIPS*, pp. 1249–1256, 2007.

[41] M. Kim, "Large margin cost-sensitive learning of conditional random fields," *Pattern Recognition*, vol. 43, no. 10, pp. 3683 – 3692, 2010.

[42] K. Prkachin and P. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, 2008.

[43] S. Lucey, A. B. Ashraf, and J. Cohn, "Investigating spontaneous facial action recognition through aam representations of the face," *Face Recognition Book*, 2007.

[44] S. Chew, S. Lucey, P. Lucey, S. Sridharan, and J. Cohn, "Improved facial expression recognition via uni-hyperplane classification," in *IEEE CVPR*, 2012, pp. 2554–2561.

[45] P. E. Shorut and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[46] J. S. Cardoso and R. Sousa, "Measuring the performance of ordinal classification," *Int'l Journ. of Pattern Recognition and Artificial Intell.*, vol. 25, no. 8, pp. 1173–1195, 2011.

[47] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *JMLR*, vol. 7, pp. 1–30, Dec. 2006.

[48] A. Agresti, *Analysis of ordinal categorical data*. Wiley Series in Prob. and Stat., 1984.

[49] Z. Zhang, M. Kim, F. De la Torre, and W. Zhang, "A real-time system for head tracking and pose estimation," in *Trends and Topics in Computer Vision*, 2012, vol. 6553, pp. 329–341.

[50] T. Yan and Q. Ji, "Automatic eye position detection and tracking under natural facial movement," in *Passive Eye Monitoring*, 2008, pp. 83–107.

Ognjen Rudovic received a BSc degree in Automatic Control from Faculty of Electrical Engineering, University of Belgrade, Serbia, in 2007, and a MSc degree in Computer Vision from Computer Vision Center (CVC), Universitat Autonoma de Barcelona, Spain, in 2008. He is currently working toward a PhD degree in 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.