Context-sensitive Conditional Ordinal Random Fields for Facial Action Intensity Estimation

Ognjen Rudovic[1], Vladimir Pavlovic[2] and Maja Pantic[1,3]
[1] Computing Department, Imperial College London, UK
[2] Department of Computer Science, Rutgers University, USA
[3] EEMCS, University of Twente, The Netherlands
{o.rudovic,m.pantic}@imperial.ac.uk http://ibug.doc.ic.ac.uk
vladimir@cs.rutgers.edu http://seqam.rutgers.edu

Abstract

We address the problem of modeling intensity levels of facial actions in video sequences. The intensity sequences often exhibit a large variability due to the context factors, such as the person-specific facial expressiveness or changes in illumination. Existing methods usually attempt to normalize this variability in data using different feature-selection and/or data pre-processing schemes. Consequently, they ignore the context in which the target facial actions occur. We propose a novel Conditional Random Field (CRF) based ordinal 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. In particular, we focus on three contextual questions: Who (the observed person), How (the changes in facial expressions), and When (the timing of the facial expression intensity). The contextual questions Who and How are modeled by means of the newly introduced covariate effects, while the contextual question When is modeled in terms of temporal correlation between the intensity levels. We also introduce a weighted softmax-margin learning of CRFs from the data with a skewed distribution of the intensity levels, as commonly encountered in spontaneous facial data. The proposed model is evaluated for intensity estimation of facial action units and facial expressions of pain from the UNBC Shoulder Pain dataset. Our experimental results show the effectiveness of the proposed approach.

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 [7]. They communicate emotions, clarify and stress what is being said, and signal comprehension, disagreement and intentions. Machine understanding of facial expressions could revolutionise user interfaces for artifacts such as robots, mobile devices, cars, and conversational agents [21], and has therefore become a hot topic in Computer Vision and Machine Learning community.

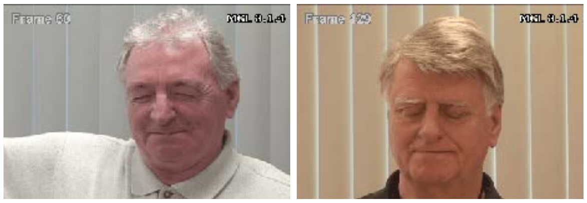

(a) AU6C, PSPI=6 (b) AU6C, PSPI=6

Figure 1. Example images of two subjects from the UNBC Shoulder Pain dataset [17], whose facial action unit AU6 (cheek raiser and lid compressor) was coded with intensity C on the A-B-C-D-E ordinal scale. The intensity of pain was computed from the codes of the co-occurring AUs in the images shown, using Prkachin and Solomon Pain Intensity (PSPI) rating. Observe the difference in the facial appearance of these two subjects whose AU6 and expressions of pain have the same intensity.

A common task in analyzing video sequences of human facial actions is to divide the sequence into segments corresponding to different phases or intensities of the target action. For example, the expression of different facial actions follow an envelope of neutral-increase-peak-decrease. Modeling such an envelop is critical for faithful representation of sequences and, consequently, for their accurate automated tagging or classification. However, in many domains and, in particular, in analysis of human affect, tagging of intensities is a burdensome task. For instance, the Facial Action Coding System (FACS) [9] defines 32 atomic facial muscle actions named Action Units (AUs), where the intensity of each AU ranges from being absent to having maximal intensity on a six-point ordinal scale. The coding of an AU intensity is carried out by a trained human FACS coder. Nevertheless, this process is tedious and error prone [18]. For that reason, most work on automatic analysis of facial actions to date has focused on detection of presence/ absence of facial actions instead of their full range intensity estimation. This is also true because the intensity reflects variability in person-specific facial expressiveness (see Fig.1), head-movements, illumination changes, and, to some extent, the annotator bias, all of which make the target problem highly *context* sensitive.

In this paper, we propose a novel context-sensitive conditional ordinal random field (cs-CORF) model for estimation of sequences of ordinal intensity of facial actions. We base our model on the general framework of CRFs [16] because of its ability to directly predict the labeling given a sequence of measurement features and the ease with which the arbitrary functions of the observed features can be incorporated into the training process. Specifically, we generalize the ordinal CRF models (CORFs) proposed in [15, 25], where the node features were designed using the homoscedastic ordinal probit model [30]. A key difference in our approach is that we account for the omnipresent impact of context on the intensity estimation by modeling the context-sensitive variability in data. To this end, we adopt the widely accepted *W5+* context model [22], where six questions are used to summarize the key aspects of the context in which the target facial action occurs: 'who' (observed person, identity, age and facial expressiveness), 'where' (environmental characteristics such as illumination), 'what' (task-specific observations of 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 funny videos).

The previously proposed approaches to AU intensity estimation (e.g., [18, 24, 26, 15, 25]) focus mainly on answering the context question 'how' by means of the covariates obtained, for instance, from expressive images after the personal texture normalization is applied. Thus, these approaches are context-free since they do not answer the other context questions. By contrast, in our approach we model the context by answering the following context questions: 'who', 'how' and 'when'. The context questions 'who' and 'how' are modeled by introducing separate context covariate effects, named CCE, and context-free covariate effects, named FCE (which coincide with the covariates used in the context-free models), respectively. These effects are efficiently embedded in the ordinal probit function, used to define the node features of the cs-CORF model. Likewise, the context question 'when' is modeled by the state-transition process used to define the edge features in the model. The CCE component is of particular importance since it directly accounts for the person-specific bias in the model's parameters, which is induced from the subject's characteristics that are considered constant across a sequence (e.g., the neutral face of the target person). We also account for heteroscedasticity in the ordinal model by allowing the model's variance to change, depending on both the CCE and FCE components. This further enhances the capacity of the model to adapt to the facial expressiveness of each person. Lastly, to address the problem of the label 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 [29].

2. Related Work

Within the context of facial affect, only a few works addressed the problem of intensity estimation of facial actions. Mahoor et al. [18] applied Spectral Regression to AAM-normalized facial appearance of infants to find the AU-specific subspaces, where the SVM classification of the intensities was then performed. In their more recent work [19], the authors applied the same approach, but to different input features, to perform intensity estimation of AUs of people watching funny videos. For the task of pain intensity estimation, Hammal and Cohn [10] proposed facial descriptors based on log-linear filters that were classified into four different levels using SVMs. The other group of methods treats the intensity levels as continuous variables. For instance, Savran et al. [26] used SVM scores obtained from the AU detectors trained on various 2D/3D image descriptors to perform continuous AU intensity estimation. Likewise, Kaltwang et al. [14] proposed a three-step approach for pain and AU intensity estimation based on fusion of different shape and appearance features using relevance vector regression (RVR). Also, Jeni et al. [11] proposed a sparse representation of facial appearance obtained by applying personal mean texture normalization to image patches. The resulting features were then used as input to the SVR model trained for intensity estimation of AUs.

All the works mentioned above focus on feature extraction, while the classification/regression is attained by using SVM/RVR, resulting in mismatch between the ordinal nature of the data and the modeling framework. In addition, skewed distribution of the intensity data poses additional difficulty to these models when learning minority classes (i.e., the higher intensities). Most importantly, none of those works addressed the omnipresent impact of context on the facial action intensity estimation. Note that context modeling has been addressed in other domains such as image annotation (e.g., [12, 31]) or activity recognition (e.g., [28, 32]). These approaches typically model the context in terms of co-occurrences of different classes (objects or activities) using CRFs for *nominal* data. By contrast, we perform *ordinal* modeling of sequences, which is preferred when dealing with intensity of actions, and 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 the intensity estimation of spontaneous facial actions.

3. Ordinal Regression

To deal with ordinal responses, different ordinal regression models have been proposed (see [5] for an overview). We restrict our consideration to the threshold model with the probit link function proposed by McCullagh(1980) [20]. In this model, the cumulative probits λ_k for $k = 1 \dots K$ ordinal responses are defined as:

$$\lambda_k = \frac{\gamma_k - \beta^T x}{\sigma}, \tag{1}$$

where x is the $D \times 1$ covariate vector, β is a vector of regression parameters, and $\gamma_0 = -\infty \leq \cdots \leq \gamma_K = \infty$ are the thresholds or cut-off points, enforcing the ordinal constraints. The scale of the cumulative probits is usually set as $\sigma \equiv 1$ for identification purpose [30]. The conditional probability of the ordinal response y is then given by:

$$\Pr(y = k|x) = \Phi(\lambda_k) - \Phi(\lambda_{k-1}), \tag{2}$$

where $\Phi(\cdot)$ represents the normal cumulative distribution function (cdf).

The most critical aspect that differentiates the ordinal regression from the multi-class classification is the modeling strategy: while the former learns a single projection (β), thus, having the same effect on the covariates across all ordinal responses, the latter learns separate hyper-planes for each class (β_k, $k = 1, ..., K$) [30]. To separate K ordinal categories, the ordinal model uses the single projection β and the cut-off points γ_k. If the responses are indeed of ordinal nature, this model is more parsimonious and often more robust than its nominal counterpart [30].

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

In this section, we first introduce the concept of context sensitive modeling of intensity levels. We then extend 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 dependence between the outputs. 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 regularizers used and the inference procedure.

4.1. Context-sensitive modeling

The context-sensitive modeling of the data is attained by allowing different effects, corresponding to different context questions, to influence the output responses y via the cumulative probit function defined in (1). We demonstrate this on the context questions 'who' and 'how', however, addressing the other context questions can be done in a similar manner. To this end, we introduce separate context covariate effects (CCE) and context-free covariate effects (FCE), which relate to the context questions 'who' and 'how', respectively. In this work, the latter are referred to as the context-free covariates as they coincide with the covariates used in traditional context-free models (e.g.,[15]). For the target task, i.e., AU intensity estimation, we define the CCE and FCE as follows. Given a sequence, $\mathbf{y}_i = \{y_{i1}, \dots, y_{iT_i}\}$, with covariates $\mathbf{x}_i = \{x_{i1}, \dots, x_{iT_i}\}$, we decompose the covariate x_{ij} into CCE ($x_i^u = C^{-1} \sum_{c=1}^{C} x_{ic}$) and FCE ($x_{ij}^r = x_{ij} - x_i^u$) components. The CCE component is ascribed to person identity (e.g., age and gender), and is computed from the first C neutral frames in an image sequence[1]. On the other hand, the FCE component accounts for variability in the facial action intensity *within* a particular sequence. We use these newly introduced effects to define the context-sensitive cumulative probits as

$$\lambda_{ijk} = \gamma_k - \beta_u^T x_i^u - \beta_r^T x_{ij}^r,\ k = 1, \dots K, \tag{3}$$

where $\sigma = 1$. From (3), we can distinguish between (i) an overall effect of the CCE component on K responses, as measured by the association of x_i^u with the responses across the whole sequence, and (ii) the effects of the FCE component on each particular response within the sequence. In other words, the CCE component is constant across the sequence, while the FCE component is time-varying. The role of the CCE component in the model can easily be seen from (3): it adjusts the locations of the thresholds γ_k in the cumulative probits depending on the target person. Thus, the simultaneous interaction of the CCE and FCE components with the other parameters of the model is what constitutes the context here.

4.2. Variance modeling

In Sec.4.3, the homoscedastic ordinal probit model is used, i.e., its variance σ^2 is assumed constant. However, since the CCE 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. For the model to be able to fully adapt to expressiveness levels of different persons, we also need to allow the scale of the thresholds to change. This can be attained

[1] We average the first five neutral frames in a sequence to obtain a more robust estimate of the target effects, however, a single frame should suffice.

by relaxing the assumption of constant σ, i.e., by allowing the noise level to vary as a function of the covariates. The ordinal models with varying noise levels are usually termed heteroscedastic ordinal models [30]. Formally, we define independent Gaussian noise terms for the CCE and FCE components, resulting in the distribution of the overall noise in the model being a zero-mean Gaussian with the variance

$$\sigma^2(x_{ij}) = \sigma_u^2(x_i^u) + \sigma_r^2(x_{ij}^r) + \sigma_o^2, \tag{4}$$

The first two terms on the right represent the CCE and FCE 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 degree of influence of the CCE and FCE variances, respectively, and log function ensures that the standard deviation is positive. We also keep the constant noise term (σ_o^2) to account for sources of variation that are not included in the model (e.g., the effects of the other context questions). The context-sensitive cumulative probits, that also have the changing variance, are now defined as

$$\lambda_{ijk} = \gamma_k \sigma^{-1}(x_{ij}) - (\beta_u^T x_i^u + \beta_r^T x_{ij}^r)\sigma^{-1}(x_{ij}), \tag{5}$$

which are used to obtain the probability of the ordinal outputs as $P(y_{ij} = k|\mathbf{x}_i) = \Phi(\lambda_{ij,k}) - \Phi(\lambda_{ij,k-1})$. From (5), we see that both the constant CCE and time-varying FCE covariates influence the scale of the model's thresholds as well as its location, thus, allowing it to adapt to the context above and beyond the contribution of the CCE effects.

4.3. Temporal modeling of ordinal data

In this section, we address the context question 'when' by encoding temporal relations between the responses. For this, we employ the linear-chain CRF [16] model that represents the conditional distribution $P(\mathbf{y}_i|\mathbf{x}_i)$, $i = 1, \dots, N$, as the Gibbs form clamped on the observations $\mathbf{x}_i$:

$$P(\mathbf{y}_i|\mathbf{x}_i;\theta) = \frac{\exp(\sum_{j=1}^{T_i} \Psi(y_{i,j-1}, y_{ij}, x_i; \theta))}{\sum\limits_{\bar{\mathbf{y}} \in \mathcal{Y}^{|T_i|}} \exp(\sum_{j=1}^{T_i} \Psi(\bar{y}_{i,j-1}, \bar{y}_{ij}, x_i; \theta))}, \tag{6}$$

where T_i is the duration of the i-th sequence, and $\mathcal{Y}^{|T_i|}$ is the set of all possible output configurations of an output graph $G = (V, E)$. $\boldsymbol{\theta}$ are the parameters of the score function $\Psi(y_{i,j-1}, y_{ij}, x_i; \boldsymbol{\theta}) \equiv \Psi_{ij}(y)$[2] defined on *node* cliques ($r \in V$) and *edge* cliques ($e = (s, r) \in E$) of the graph as

$$\Psi_{ij}(y) = f_n(y_{ij}, x_i) + f_e(y_{i,j-1}, y_{ij}). \tag{7}$$

The choice of the *node* $f_n(y_{ij}, x_i)$ 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 cumulative probits to set the *node* features as

$$f_n(y_{ij}, \mathbf{x}_i) = \sum_{k=1}^{K} I(y_{ij} = k) \cdot \log P(y_{ij} = k|\mathbf{x}_i), \tag{8}$$

where $P(y_{ij} = k|\mathbf{x}_i) = \Phi(\lambda_{ij,k}) - \Phi(\lambda_{ij,k-1})$, and $I(\cdot)$ is the indicator function that returns 1 (0) if the argument is true (false). The *edge* features model the first order Markov dependence between the ordinal responses as

$$f_e(y_{i,j-1}, y_{ij}) = \sum_{m,k=1}^{K} I(y_{i,j-1} = m \wedge \, y_{ij} = k) \cdot u_{mk}, \tag{9}$$

where $m, k{=}1 \dots K$, and u_{mk} measures the temporal association between the responses. Note that the denominator of (6) guarantees that the distribution sums to one, and is computed using (8, 9), but without the indicator function. Now, given the iid training data $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^N$, the 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*

$$\min_\theta R(\theta) - \sum\nolimits_{i=1}^N \log P(\mathbf{y}_i|\mathbf{x}_i;\theta), \tag{10}$$

where $R(\theta)$ is the regularization term that prevents the model from overfitting.

We term this model the Context-sensitive Conditional Ordinal Random Field (cs-CORF) model since it uses the newly proposed context-sensitive cumulative probits in its design of the *node* features. Note that the model in [15] is a special case of the cs-CORF model, and it can be obtained by removing the context from the cumulative probits in (5), and by setting its variance constant. However, as we show in our experiments, these effects play a crucial role in raising the estimation performance.

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 [27] to the CRF model in (6). However, in contrast to [27], we introduce scaling of the slack variables, which incurs a higher penalty when making errors on minority classes during learning. We start from the standard primal learning approach for max-margin models [29, 13]:

$$\begin{aligned} &\min_{\zeta_{ij},\theta} R(\theta) + \sum\nolimits_{i=1}^N \sum\nolimits_{j=1}^{T_i} \zeta_{ij} \\ &s.t.\; \Psi_{ij}(y) - \Psi_{ij}(\bar{y}) \geq \Delta_{ij}(y, \bar{y}) - \tfrac{\zeta_{ij}}{w_{ij}(y,\bar{y})}, \\ &\forall \bar{y}_{ij} \in \mathcal{Y}, \zeta_{ij} > 0,\, i = 1 \dots N, j = 1 \dots T_i, \end{aligned} \tag{11}$$

where the large-margin set of constraints are applied to the score function defined in (7). These constraints enforce the difference between the scores of the correctly

[2] For notational simplicity, we drop dependence on $j-1$, x_i and θ.

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 temporally neighboring pairs of labels as the weighted Hamming loss, i.e., $\Delta_{ij}(y, \bar{y}) = 1 - [\alpha I(y_{ij}, \bar{y}_{ij}) + (1-\alpha) I(y_{ij-1}, \bar{y}_{ij-1})]$, for $j{>}1$ and $0 \leq \alpha \leq 1$, while for the first example in the sequence ($j{=}1$), we set $\alpha{=}1$. The weighting of the slack variables ζ_{ij} is attained using the information about prior distribution of the intensity levels as $p(y) = N_y / \sum_{k=1}^{K} N_k$, leading to $w_{ij}(y, \bar{y}) = w_{ij}(y) = 1/(p(y) + \varepsilon)$. The parameter ε is chosen from the range $[0, 1]$, and it ensures that the overall loss is not dominated by minority classes. The constraint in (11) can be written as

$$w_{ij}(y)\Psi_{ij}(y) - w_{ij}(y)(\Psi_{ij}(\bar{y}) + \Delta_{ij}(y, \bar{y})) \geq -\zeta_{ij}, \quad (12)$$

Note that when the weight $w_{ij}(y)$ is set to one, the constraint in (12) is equivalent to that used in the conventional n-Slack large-margin learning with margin-rescaling (e.g., [29]). If this constraint is satisfied for each clique, then it will be satisfied for the whole graph in the CRF. So, instead, we will require that:

$$\begin{aligned} &\min_{\zeta_i, \theta} R(\theta) + \textstyle\sum_{i=1}^{N} \zeta_i \\ &s.t. \ \textstyle\sum_{j=1}^{T_i} \left[\Psi_{ij}^w(y) - (\Psi_{ij}^w(\bar{y}) + \Delta_{ij}^w(y, \bar{y}))\right] \geq -\zeta_i, \\ &\forall \bar{y}_{ij} \in \mathcal{Y}^{|T_i|}, \ i = 1 \ldots N, \ \zeta_i > 0, \end{aligned} \quad (13)$$

where we simplify 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 $\Delta_{ij}^w(y, \bar{y}) \equiv w_{ij}(y)\Delta_{ij}(y, \bar{y})$.

Next, for given θ, each ζ_i in the optimization problem (OP) in (13) can be optimized individually [13], and the smallest feasible ζ_i, given θ, is achieved for:

$$\zeta_i = \max_{\bar{\mathbf{y}}_\mathbf{i} \in \mathcal{Y}^{|T_i|}} \sum_{j=1}^{T_i} (\Psi_{ij}^w(\bar{y}) + \Delta_{ij}^w(y, \bar{y})) - \sum_{j=1}^{T_i} \Psi_{ij}^w(y) \quad (14)$$

We now 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

$$\zeta_i = \log \sum_{\bar{\mathbf{y}}_\mathbf{i} \in \mathcal{Y}^{|T_i|}} \mathrm{e}^{\sum_{j=1}^{T_i} \Psi_{ij}^w(\bar{y}) + \Delta_{ij}^w(y, \bar{y})} - \sum_{j=1}^{T_i} \Psi_{ij}^w(y) \quad (15)$$

The constraint in (15) is more restrictive than that in (14) 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 (13) as an unconstrained OP. Specifically, since the constraint in (15) has a form similar to that of the negative log of the conditional probability of CRFs defined in (6), we can formulate the weighted softmax-margin learning of the CRF/cs-CORF model as the following (unconstrained) OP:

$$\min_{\zeta_i, \theta} R(\theta) + \sum_{i=1}^{N} \zeta_i \equiv \min_{\theta} R(\theta) - \sum_{i=1}^{N} \log P^w(\mathbf{y}_i | \mathbf{x}_i; \theta), \quad (16)$$

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

$$P^w(\mathbf{y}_i | \mathbf{x}_i; \theta) = \frac{\exp(\sum_{j=1}^{T_i} \Psi_{ij}^w(y))}{\sum\limits_{\bar{\mathbf{y}} \in \mathcal{Y}^{|T_i|}} \exp(\sum_{j=1}^{T_i} \Psi_{ij}^w(\bar{y}) + \Delta_{ij}^w(y, \bar{y}))} \quad (17)$$

The introduced formulation of the weighted softmax large-margin learning allows us to compute the model parameters θ efficiently by using the gradient optimization and dynamic programming techniques (e.g., Viterbi algorithm), commonly employed for learning of CRFs. Thus, the implementation is straightforward as it only requires applying the weights to the score function $\Psi(\cdot)$ penalized with the loss $\Delta(\cdot)$. On the other hand, inference is performed by using the unweighted/unpenalized likelihood in (6).

4.5. Regulizers

To deal with the order constraints in the parameters γ, we introduce the displacement variables δ_k, where $\gamma_j = \gamma_1 + \sum_{k=1}^{j-1} \delta_k^2$ for $j = 2, \ldots, K-1$. So, γ is replaced by the unconstrained parameters $\{\gamma_1, \delta_1, \ldots, \delta_{K-2}\}$. Another important issue is the regularization of the parameters of the cs-CORF model. We use the L_2 regularizer for the standard CRF parameters, resulting in the regularization term $R(\theta)$ defined as:

$$R(\theta) = \rho_1(\|\beta_u\|^2 + \|v_u\|^2) + \rho_2(\|\beta_r\|^2 + \|v_r\|^2) + \rho_3\|u\|^2, \quad (18)$$

where (ρ_1, ρ_2, ρ_3) are the constants controlling the penalty of the node and edge potentials, respectively. Furthermore, with ρ_1 and ρ_2 we allow the model to adequately balance the impact of the CCE and FCE effects. With $R(\theta)$, as defined in (18), the objective in (16) can be minimized by applying any unconstrained optimizer. Here we use the quasi-Newton LBFGS method. The regularization parameters are found using a validation procedure on the training set. Once the model parameters are estimated, the inference of test sequences is carried out by applying Viterbi decoding to the 'unweighted' conditional likelihood in (6).

5. Experiments

Evaluation of the proposed model is performed using the UNBC-MacMaster Shoulder Pain Expression Archive (Shoulder-Pain) dataset [17] containing video recordings of patients suffering from shoulder pain while performing range-of-motion tests of their arms. 200 sequences

		SVM	GPOR	SVORIM	RVM	SR+SVM	CRF(ml)	CRF(w)	CORF(ml)	CORF(w)	CORF(ml+h)	CORF(w+h)
F-1	CCE+FCE	23.1 (18.2)	24.7 (17.6)	28.5 (14.2)	27.9 (17.2)	32.1 (13.5)	34.2 (11.2)	36.5 (8.0)	38.3 (5.8)	40.3 (4.1)	40.5 (4.0)	**43.5 (1.6)**
	FCE	23.5 (17.7)	23.2 (18.5)	25.2 (16.1)	29.9 (15.9)	34.4 (10.2)	33.5 (11.5)	35.0 (9.7)	34.3 (11.1)	36.6 (8.0)	36.4 (8.7)	39.1 (5.1)
MAE	CCE+FCE	1.13 (18.9)	1.03 (17.7)	0.95 (14.7)	0.99 (16.3)	1.00 (15.3)	0.89 (11.7)	0.86 (9.3)	0.80 (6.4)	0.75 (3.2)	0.75 (3.3)	**0.69 (1.0)**
	FCE	1.16 (19.5)	1.10 (18.2)	0.99 (15.1)	1.02 (17.6)	0.88 (11.4)	0.91 (12.6)	0.88 (11.9)	0.84 (9.1)	0.81 (7.1)	0.82 (8.0)	0.78 (4.9)
ICC	CCE+FCE	32.8 (17.8)	36.7 (16.2)	38.4 (13.5)	16.9 (21.1)	29.4 (17.5)	46.0 (11.5)	51.1 (7.9)	55.1 (5.2)	57.7 (3.5)	58.2 (3.0)	**61.9 (1.2)**
	FCE	34.5 (17.1)	35.5 (16.8)	36.3 (16.5)	26.5 (19.2)	39.0 (15.1)	46.8 (11.3)	49.5 (9.3)	50.5 (8.7)	51.6 (8.1)	52.4 (7.4)	55.3 (5.2)

Table 1. The average performance of the models tested on 11 intensity estimation problems (expression of *pain* + 10 AUs from the Shoulder-pain dataset). The numbers in brackets are the average ranks of the models, where the ranking is performed on 22 (=11×2) tasks, as each model is tested using two sets of covariates: the context (CCE+FCE) and context-free (FCE) covariates. Note that for all three evaluation scores, the top ranked model is the proposed context-sensitive CORF(w+h) (i.e., CORF(w+h) with the CCE+FCE covariates).

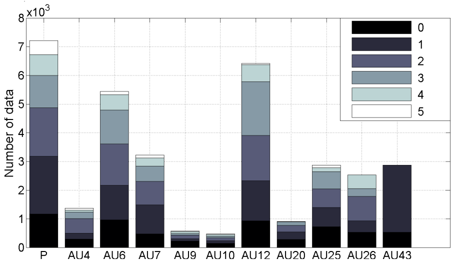


Figure 2. Distribution of the intensity levels in pre-segmented sequences of pain and AUs from the Shoulder-pain dataset.

of 25 subjects were recorded. For each frame, the FACS coding of the intensity on a 6 level ordinal scale (neutral<A<B<C<D<E) for 11 AUs is provided by the database creators. In our experiments, we model intensity of AUs 4, 6, 7, 9, 10, 12, 20, 25. Also, we used discrete pain intensities (0-15) defined according to the Prkachin and Solomon [23] measure. However, since the higher pain intensity contains only a few examples, we discretized it into 6 levels as: 0, 1, 2 ,3, 4-5, 6-15. We further pre-segmented all the image sequences containing intensity of the target AUs/pain $>$ 0, so that the number of neutral frames was balanced with the second most frequent intensity. Still, the resulting intensity distribution remained skewed toward low levels (see Fig. 2). As input features, we used the locations of 66 facial landmark points extracted by the database creators using an Active Appearance Model (AAM) [17]. To reduce the effect of head movements, these points were registered to a reference face using an affine transform. Lastly, to reduce the feature dimensionality, we applied PCA to data of each AU, resulting in 18-D feature vectors, on average, where 97% of energy was preserved. The CCE and FCE covariates were then obtained as explained in Sec.4.3.

We compare the performance of the context-sensitive and context-free CORF model, as well as 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 ($\sigma = 1$) and heteroscedastic noise models ($\sigma(x)$), with the latter denoted by '**h**'. We also show the performance of the standard linear-chain CRF model [16], trained using both 'ml' and 'w' learning. As the baseline model, we use one-vs-all SVM [4]. We also perform comparisons with the state-of-the-art *static* ordinal regression models, Support Vector Ordinal Regression with implicit constraints (SVORIM) [6] and Gaussian Process Ordinal Regression [5]. In the kernel methods (SVM/SVOR/GPOR), we used the linear kernel. Finally, we include the comparisons with the state-of-the-art models for AU intensity estimation: the RVM approach [14] for continuous estimation of AU intensity, and Spectral Regression [2] combined with one-vs-one SVM (SR+SVM) [18, 19]. 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 the methods we used the RBF kernel, as in [14, 19]. The hyper/regularization-parameters of all methods were selected by a validation on the training set. For testing, we applied a 5-fold cross validation procedure, with each fold containing intensity sequences of different persons. We report the accuracy of the models using: the average of F-1 scores computed for each intensity level, the weighted mean absolute error (MAE) [1], and Intra-Class Correlation (ICC), commonly used in behavioral sciences to quantify agreement between (human) coders (see [18] for details).

Table 1 shows the average results obtained by different models on 11 intensity estimation problems (*pain* + 10 AUs). The models were trained/tested using two sets of covariates: context (CCE+FCE) and context-free (FCE). To ensure that the models' performance is consistent across most of the tasks, we computed average rankings of the models across all 22 tasks (11 problems × 2 sets of covariates). The models are ranked for each task separately, 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 ranks were obtained by averaging the rankings over all tasks, as in [8] c.f. Sec.3.2.2.

We see from Table 1 that inclusion of the CCE component in traditional static approaches does not necessarily improve their performance. We found that these models were very sensitive to overfitting of the CCE component, regardless of the regularization employed. This is especially pronounced in the SR+SVM model, where the SR-learned subspaces were biased towards the training subjects. The

Shoulder-Pain		P	AU4	AU6	AU7	AU9	AU10	AU12	AU20	AU25	AU26	AU43
F-1	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**
	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
	CRF(w)	30.0	27.0	29.0	29.0	33.0	42.0	32.0	32.0	29.0	26.0	76.0
	RVM	22.8	26.7	22.2	22.1	23.5	43.0	27.8	25.5	22.1	22.0	70.7
	SR+SVM	29.4	24.3	23.9	22.3	32.6	43.4	26.7	29.6	36.0	32.4	78.3
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**
	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
	CRF(w)	1.16	0.99	0.98	1.00	0.82	0.53	0.94	0.93	0.99	1.23	0.13
	RVM	1.00	1.05	1.16	1.25	1.30	0.64	0.98	0.99	1.16	1.50	0.18
	SR+SVM	1.00	0.93	0.97	1.13	0.85	0.63	0.81	0.85	0.97	1.39	0.11
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**
	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
	CRF(w)	58.0	66.0	52.0	54.0	52.0	49.0	51.0	37.0	43.0	29.0	54.0
	RVM	43.1	33.9	18.8	28.9	-0.5	39.1	27.7	16.3	21.7	16.8	46.0
	SR+SVM	44.4	54.6	36.0	27.2	43.4	37.8	34.0	35.2	38.8	18.2	59.1

Table 2. The performance of the models on intensity estimation of expression of *pain* (P) and 10 AUs from the Shoulder-pain dataset. The numbers in *bold* indicate that the proposed cs-CORF(w+h) performs significantly better than the rest of the models, based on the paired t-test with $p = 0.05$.

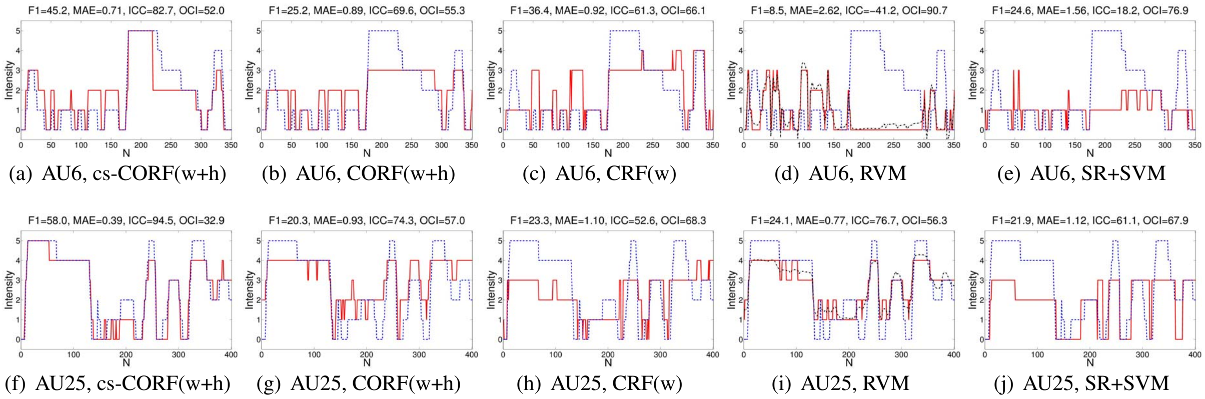
static ordinal models, GPOR and SVOR, showed a small improvement in their performance when the context covariates are used. Furthermore, SVOR performed better than the static nominal models in terms of MAE and ICC, both of which are better suited for measuring ordinal performance than F-1. On the other hand, the temporal models (CRFs and CORFs) significantly increase the performance of the static methods. This is even more pronounced when the proposed weighted soft-max learning is used. Moreover, we see that the parameter tying in the CORF models, especially in the presence of the context covariates, results in their overall better performance over CRFs, with each introduced effect enhancing the evaluation scores. Based on the ranking of the models, the proposed cs-CORF(w+h) consistently outperforms the other models in most of the tasks.

Table 2 shows the performance of different models in each task. Here we compare cs-CORF(w+h) with context-free CORF(w+h) and CRF(w). We also include the results obtained by the context-free SR+SVM[19] and RVM[14] models, which have previously been used for the target task. The numbers in bold in Table 2 indicate that the differences in the 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). Again, the proposed cs-CORF(w+h) model performs similarly or better than the context-free models in most tasks. Note, for instance, that since AU10 involves activation of vertically set muscles above the upper lip, no strong personal characterization is expected. Thus, modeling of the context question 'who' in cs-CORF does not much improve the performance of the context-free CORF model. This is in contrast to AU12, which involves activation of an oblique muscle, resulting in curved facial motion that varies considerably across persons.

Fig.3 shows the intensity estimation at the sequence level of two example AUs, namely AU6 and AU25. The scores shown in the title of each graph are computed from the depicted sequences. Here we also include the Ordinal Classification Index (OCI) [3] score, whose lower values indicate less confusion among the neighboring levels. We see that the RVM model estimates well the slope of the true signal, but it misses its scale, which is a consequence of assuming an equal interval scale for the outputs. On the other hand, SR+SVM underestimates the true intensity levels possibly because of its bias toward the majority classes in the learned subspace. Based on the obtained scores, CRF(w) performs better than CORF(w+h) in terms of F-1, while the latter model achieves better MAE and ICC, which are better suited for ordinal data. However, cs-CORF(w+h) outperforms the other models in all aspects, especially in the case of higher intensity levels. We attribute this to its ability to answer the context question 'who', in addition to 'how' and 'when', and, therefore, properly adapt to the facial expressiveness of different persons.

6. Discussion and Conclusions

The results obtained indicate the benefits of each proposed improvement. We show that the inclusion of the context question 'who' is critical for substantially raising the performance of standard CORF(ml) across all three scoring measures. Traditional static models do not account for impact of the context on output responses in a principled manner, which evidently limits their estimation performance. As we show in the experiments, because of lack of parameter tying as well as sequential modeling these models fail to fully exploit the context component (CCE). While the CRF nominal model performs well (with the inclusion of CCE + FCE), it fails to reach the full performance level of cs-CORF. This is because of the lack of the ordering constraints and, possibly, because of the increased parameter dimensionality. Finally, due to the unbalanced nature of our data, a proper scaling of the loss during training is crucial. The most frequent low intensity levels that would otherwise dominate performance scores are properly balanced using the proposed margin balancing. This is reflected in improvements of the weighted models (w) over their unweighted counterparts (ml).



(a) AU6, cs-CORF(w+h) (b) AU6, CORF(w+h) (c) AU6, CRF(w) (d) AU6, RVM (e) AU6, SR+SVM

(f) AU25, cs-CORF(w+h) (g) AU25, CORF(w+h) (h) AU25, CRF(w) (i) AU25, RVM (j) AU25, SR+SVM

Figure 3. The ground-truth (*dashed blue*) and estimation (*solid red*) of intensity of several exemplary sequences of AUs 6 and 25. For RVM, we also include the continuous estimation of intensities (*dashed black*).

To conclude, in this paper we have proposed a simple but principled approach for context-sensitive intensity estimation of facial action units and expressions of pain. We incorporated the context in our model by answering the questions 'who', 'how' and 'when' from the W5+ context design. As shown by our experiments, the proposed model outperforms traditional models for nominal/ordinal classification, and the state-of-the-art models for intensity estimation of facial actions.

Acknowledgments

This work has been supported by the European Research Council under the ERC Starting Grant no. ERC-2007-StG-203143 (MAHNOB), and the National Science Foundation under Grant no. IIS0916812. The current work by Ognjen Rudovic is supported by the EPSRC grant EP/J017787/1: Analysis of Facial Behaviour for Security in 4D (4D-FAB).

References

[1] S. Baccianella, A. Esuli, and F. Sebastiani. Evaluation measures for ordinal regression. *Int'l Conf. on Intell. Syst. Design and Applications*, pages 283–287, 2009. 6

[2] D. Cai, X. He, and J. Han. Spectral regression for efficient regularized subspace learning. *IEEE ICCV*, pages 1–8, 2007. 6

[3] J. S. Cardoso and R. Sousa. Measuring the performance of ordinal classification. *Int'l Journ. of Pattern Recognition and Artificial Intell.*, 25(8):1173–1195, 2011. 7

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intell. Syst. and Techn.*, 2:27:1–27:27, 2011. 6

[5] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *JMLR*, 6:1019–1041, 2005. 3, 6

[6] W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. *ICML*, pages 145–152, 2005. 6

[7] 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. 1

[8] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7:1–30, Dec. 2006. 6

[9] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, 2002. 1

[10] Z. Hammal and J. F. Cohn. Automatic detection of pain intensity. *ICMI*, pages 47–52, 2012. 2

[11] 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*, pages 1–7, 2013. 2

[12] W. Jiang, S.-F. Chang, and A. C. Loui. Context-based concept fusion with boosted conditional random fields. *IEEE ICASSP*, 2007. 2

[13] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Mach. Learn.*, 77(1):27–59, 2009. 4, 5

[14] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. *ISVC*, 7432:368–377, 2012. 2, 6, 7

[15] M. Kim and V. Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. *ECCV*, pages 649–662, 2010. 2, 3, 4

[16] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, pages 282–289, 2001. 2, 4, 6

[17] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. *IEEE FG*, pages 57–64, 2011. 1, 5, 6

[18] 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*, pages 74–80, 2009. 2, 6

[19] S. Mavadati, M. Mahoor, K. Bartlett, P. Trinh, and J. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Trans. on Affective Comp.*, 4(2):151–160, 2013. 2, 6, 7

[20] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Stat. Society. Series B*, 42:109–142, 1980. 3

[21] M. Pantic. Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of Royal Society B*, 364:3505–3513, 2009. 1

[22] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. *Human Computing and Machine Understanding of Human Behavior: A Survey*, volume 4451, pages 47–71. 2007. 2

[23] K. Prkachin and P. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008. 6

[24] J. Reilly, J. Ghent, and J. McDonald. Investigating the dynamics of facial expression. *Lecture Notes in Computer Science*, 4292:334–343, 2006. 2

[25] O. Rudovic, V. Pavlovic, and M. Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. *IEEE CVPR*, pages 2634–2641, 2012. 2

[26] A. Savrana, B. Sankur, and M. Bilgeb. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 2012. 2

[27] F. Sha and L. K. Saul. Large margin hidden markov models for automatic speech recognition. *NIPS*, pages 1249–1256, 2007. 4

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

[29] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005. 2, 4, 5

[30] R. Winkelmann and S. Boes. *Analysis of microdata*. Springer, 2006. 2, 3, 4

[31] 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*, pages 3368–3375, 2010. 2

[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*, 34(9):1691–1703, 2012. 2