Bilateral Ordinal Relevance Multi-instance Regression for Facial Action Unit Intensity Estimation

Yong Zhang[1,2], Rui Zhao[3], Weiming Dong[1], Bao-Gang Hu[1], and Qiang Ji[3]*
[1]National Laboratory of Pattern Recognition, CASIA
[2]University of Chinese Academy of Sciences
[3]Rensselaer Polytechnic Institute
zhangyong201303@gmail.com, zhaor@rpi.edu, weiming.dong@ia.ac.cn,
hubg@nlpr.ia.ac.cn, qji@ecse.rpi.edu

Abstract

Automatic intensity estimation of facial action units (AUs) is challenging in two aspects. First, capturing subtle changes of facial appearance is quite difficult. Second, the annotation of AU intensity is scarce and expensive. Intensity annotation requires strong domain knowledge thus only experts are qualified. The majority of methods directly apply supervised learning techniques to AU intensity estimation while few methods exploit unlabeled samples to improve the performance. In this paper, we propose a novel weakly supervised regression model-Bilateral Ordinal Relevance Multi-instance Regression (BORMIR), which learns a frame-level intensity estimator with weakly labeled sequences. From a new perspective, we introduce relevance to model sequential data and consider two bag labels for each bag. The AU intensity estimation is formulated as a joint regressor and relevance learning problem. Temporal dynamics of both relevance and AU intensity are leveraged to build connections among labeled and unlabeled image frames to provide weak supervision. We also develop an efficient algorithm for optimization based on the alternating minimization framework. Evaluations on three expression databases demonstrate the effectiveness of the proposed method.

1. Introduction

Human facial expressions are efficient means of human communication, conveying rich information for expressing our intention and emotion. They can be described by a specific combination of facial muscle movements. Facial Action Coding System (FACS) was developed by Ekman and Friesen [8] to describe such movements. FACS defines each observable component of facial movement as an AU. It quantifies AU intensity into 6 discrete levels and provides rules for annotation. However, automatic AU intensity estimation is a challenging task. First of all, AUs depict the subtle local facial appearance changes which vary by subjects, head poses, and illuminations. Second, AU intensity annotation requires domain expertise and it is time-consuming. Only databases annotated by certificated AU coders are reliable. Thus it is expensive to annotate a large database. Although techniques like deep learning has increased modeling capacity for appearance variations, they also require a large set of labeled samples, which may not be available.

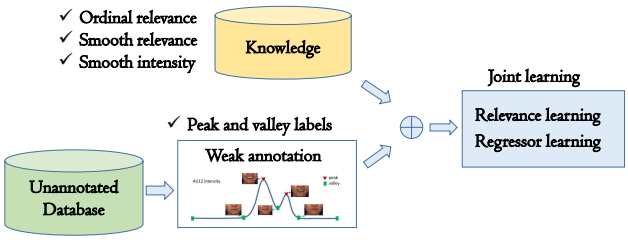

Figure 1. The pipeline of the proposed method. BORMIR combines the weakly annotated sequences and the domain knowledge to jointly learn the frame-level relevance and intensity regressor.

The majority of existing methods directly apply supervised learning techniques for AU intensity estimation, such as relevance vector regression [12] and convolutional neural networks [9]. These approaches do not consider the case when intensity annotations are limited. Few works focus on exploiting unlabeled instances to improve the performance except for [30] and [50]. They exploit unlabeled instances by considering the temporal relationships among AU intensity labels of frames in sequences.

As illustrated in Figure 2, AU intensity changes smoothly as facial appearance evolves smoothly. Instead of annotating the AU intensity for every frame in sequences, identifying the locations of the peak and valley frames (key frames) is relatively easier to perform for a large database since annotating the trend requires less effort than annotating the exact intensity. One valuable property is that the AU

*Corresponding author.

intensity between a peak and a valley frame evolves non-strictly monotonically and smoothly.

Considering such property, we propose a novel weakly supervised regression model, BORMIR, which learns a frame-level intensity regressor by using weakly labeled sequences. The weak annotation only requires identifying and annotating the peak and valley frames within a sequence. They are defined in [20]. According to the key frames, sequences can be split into segments which retain the property. For the remaining of the paper, we use bag to refer a segment and use instance to refer a frame. Unlike [30] and [50], we study AU intensity estimation from a new perspective. The annotations of the peak and valley frame in a segment are treated as bag labels. Different from conventional multi-instance learning (MIL) methods, we consider simultaneously two bag labels for each bag to include more information, i.e., *peak bag label* (the intensity of the peak) and *valley bag label* (the intensity of the valley). More importantly, we introduce the concept of *'relevance'* to model sequential data. Each instance has a relevance value to one bag label. Each bag label is contributed by all instances. Both the relevance and parameters of the regressor are the variables to be optimized during learning. To exploit unlabeled frames, we leverage domain knowledge on relevance and AU intensity. Firstly, in each segment, the closer the frame is to the peak (or valley), the larger relevance value it has to the peak (or valley) label, i.e., ordinal relevance. Secondly, the difference between the relevance of neighboring frames should be small due to the smoothness of the evolution of facial appearance, i.e., relevance smoothness. Thirdly, the difference between the intensities of neighboring frames should be small, i.e., intensity smoothness. The pipeline of the proposed method is shown in Figure 1.

Our contributions are as follows:

- We propose a novel approach to learn frame-level intensity regressor with limited annotations. We formulate the intensity estimation as a multi-instance regression problem. In particular, we introduce the concept of 'relevance' to model sequential data and simultaneously consider two bag labels for one bag to include more information.
- We leverage domain knowledge as weak supervision to make the learned regressor applicable for frame-level intensity estimation, including ordinal relevance, relevance smoothness, and intensity smoothness.
- We develop an efficient algorithm to solve the problem and evaluate the proposed method on three benchmark expression databases.

2. Related Work

Supervised methods. The majority of methods for AU intensity estimation are supervised methods. Many methods consider the intensity estimation of individual AUs such as [19, 33, 11, 10, 44, 23, 22] . Kaltwang *et al.* [12] applied Relevance Vector Regression (RVR) to estimate AU intensity. Then, they proposed the doubly sparse Relevance Vector Machine (DSRVM) [14] for intensity estimation of AU and pain expression by selecting relevance samples and important kernels. To consider the joint estimation of multiple AUs, several methods leverage static dependencies among AUs to improve the model learning such as [32, 43, 13]. Walecki *et al.* [43] proposed Copula Ordinal Regression (COR) model by using copula functions to define the pairwise potential of the conditional random field. Kaltwang *et al.* [13] proposed a generative latent tree model (LT) by learning the dependencies among both features and intensities of multiples AUs. Dynamic dependencies are commonly used to model sequential data such as action recognition [38, 24, 25] and tracking [47, 2, 49, 48]. They can also be used for AU intensity estimation. Several methods apply dynamic graphical models to perform joint AU intensity estimation, such as Dynamic Bayesian Network [16], Hidden Conditional Ordinal Random Fields (HCORF) [27, 28], context-sensitive CORF [29], and Continuous Conditional Neural Fields [1]. Deep learning has also been applied to AU intensity estimation [9, 42, 17, 7]. Gudi *et al.* [9] used convolutional neural network (CNN) for AU detection and intensity estimation in FERA 2015 [39]. Walecki *et al.* [42] combined CRF and copula functions (CCNN-IT) to jointly learn deep representation and AU relationships. Tran *et al.* [17] proposed semi-parametric variational autoencoders (2DC) for the intensity estimation of multiple AUs. Supervised methods tend to overfit the training set when intensity annotations are not sufficient, especially for deep models.

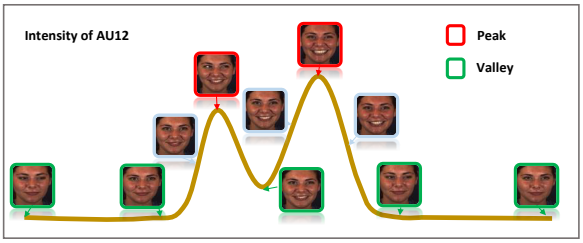

Figure 2. Illustration of AU in a sequence.

Weakly supervised methods. Multi-instance learning (MIL) provides a way to leverage weakly labeled data for model learning. It has been applied to key frame detection [37], pain localization [35] and facial event detection [31, 36]. LOMo [36] is a variant of multi-instance classification (MIC) which incorporates the ordering of indices of frames. Zhu *et al.* [52] used Bootstrapping to select positive and negative samples among frames between onset and offset to improve the performance of AU detection. Fernando *et al.* [6] leveraged unlabeled frames by directly using features to compute the similarity between one frame and the peak. No model and relevance learning are involved. However, these methods can not generalize to

AU intensity estimation since event detection is a binary classification problem while AU intensity label has 6 discrete ordinal levels. To the best of our knowledge, only two methods [30, 50] applied weakly supervised learning methods to estimate AU intensity. Zhao *et al.* [50] combined ordinal regression and SVR (OSVR) to leverage both labeled and unlabeled frames. They need to identify and annotate the intensities of the onset, apex, and offset frames. Ruiz *et al.* [30] proposed Multi-instance Dynamic Ordinal Random Fields (MI-DORF) with the idea of multi-instance learning. The maximum intensity of each sequence is annotated as the bag label. [50] and [30] leverage unlabeled samples by using the relationships among intensity labels of neighboring frames. Differently, we introduce 'relevance' to model sequential data and emphasize domain knowledge on both relevance and AU intensity to exploit unlabeled samples including ordinal relevance and relevance smoothness. And we consider two labels for one bag by treating the intensities of the peak and valley frames as bag labels.

Multi-instance regression. Few works focus on Multi-Instance Regression (MIR). Ray *et al.* [26] pioneered MIR research under the prime instance assumption that each bag has a prime instance responsible for the bag label. But it is inapplicable to unseen bag since the primary instance of an unseen bag is unknown. Wagstaff and Lane [40] proposed a new assumption that each instance has a weight for the bag label and the bag label is the weighted summation of predictions of all instances. However, the work aims to learn the salience of instances and it does not discuss how to predict the weight or bag label of an unseen instance or bag. Different assumptions vary from tasks [3, 45, 41]. But their goal is to predict the label for an unseen bag rather than each instance. Our method makes the same assumption as [40]. Unlike [40], our method considers simultaneously two bag labels for one bag and our goal is to learn an instance-level regressor while [40] aims to estimate the weight only and it performs poorly for instance-level prediction. Besides, we exploit relationships among instances and incorporate different types of domain knowledge to make the learning of instance-level regressor feasible.

3. The Proposed Method

Notation Given the locations and intensity annotations of peak and valley frames, training sequences can be split into segments according to [20]. Each segment has the property that the intensity increases or decreases monotonically and smoothly. The intensity of the peak frame is treated as the peak bag label since it is more informative. The intensity of the valley frame is treated as the valley bag label. Each frame has two relevance values corresponding to the peak and valley labels respectively, *i.e.*, *peak relevance* and *valley relevance*. Instead of introducing a variable to specify the trend of segments (increase or decrease), we rearrange the frame order for segments which start from a peak and end with a valley. After the rearrangement, the intensity in each training segment monotonically increases.

The training set $\mathcal{D} = \{(\mathbf{B}_i, y_i^0, y_i)\}_{i=1}^N$ contains of N segments. Let $\mathbf{B}_i = [\mathbf{B}_i^1, \mathbf{B}_i^2, ..., \mathbf{B}_i^{n_i}] \in \mathbb{R}^{d\times n_i}$ denote the image features of the i-th segment, where n_i is the number of frames and d is the feature dimension. The peak bag label $y_i \in \mathbb{R}$ is the intensity of the peak frame. The valley bag label $y_i^0 \in \mathbb{R}$ is the intensity of the valley frame. Let $\boldsymbol{\alpha}_i^j$ and $\boldsymbol{\beta}_i^j$ denote the peak and valley relevance of the j-th frame, then we have $\mathbf{H}_i = \sum_{j=1}^{n_i} \boldsymbol{\alpha}_i^j \mathbf{B}_i^j = \mathbf{B}_i\boldsymbol{\alpha}_i$ as the combination of instances in a bag. To force the combination locate in the convex hull of these instances, $\boldsymbol{\alpha}_i$ should satisfy $\sum_{j=1}^{n_i} \boldsymbol{\alpha}_i^j = 1$ and $\boldsymbol{\alpha}_i \geq 0$.

Given the training set $\mathcal{D}$, the goal is to learn a frame-level intensity estimator f to predict the intensity y for an unlabeled frame $\mathbf{x} \in \mathbb{R}^d$, *i.e.* $y = f(\mathbf{x}; \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^d$ are the parameters of the estimator. We use a linear model for intensity estimation, *i.e.*, $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T\mathbf{x}$. During testing, we perform frame-level intensity prediction through $y = \mathbf{w}^T\mathbf{x}$ for an unseen frame.

3.1. Peak and Valley Bag Labels

Conventional MIR methods consider only one bag label for each bag. When applying MIR for AU intensity estimation, only the intensity of the peak frame is exploited as the bag label. The drawback is that it ignores the information of frames around the valley frame which are far away from the peak. In our proposal, the intensities of both the peak and valley frames are considered, to ensure the regressor fit all frames. In particular, the valley bag label provides information around the valley frame. Similar to peak relevance $\boldsymbol{\alpha}_i$, the valley relevance $\boldsymbol{\beta}_i$ should satisfy $\sum_{j=1}^{n_i} \boldsymbol{\beta}_i^j = 1$ and $\boldsymbol{\beta}_i \geq 0$. For each frame, the peak and valley relevance are not independent. The two relevance values have such association that when $\boldsymbol{\alpha}_i^j$ is small, $\boldsymbol{\beta}_i^j$ should be relatively large. To represent the correlation, the summation of two relevance values of each frame is the same, *i.e.*, $\boldsymbol{\alpha}_i^j + \boldsymbol{\beta}_i^j = \boldsymbol{\alpha}_i^k + \boldsymbol{\beta}_i^k$ for the j and k-th frames. The equivalent representation is

$$\mathbf{V}_i(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i) = \mathbf{0}, \tag{1}$$

where $\mathbf{V}_i \in \mathbb{R}^{n_i\times n_i}$ with $\mathbf{V}_i^{j,j} = 1$, $\mathbf{V}_i^{j,j+1} = -1$, and other elements being 0. The loss for the peak bag label is

$$L(\mathbf{w}, \{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{B}_i\boldsymbol{\alpha}_i)^2. \tag{2}$$

It can be interpreted as the square loss between the ground truth bag label and the predicted bag label of the average instance. Similarly, the loss for the valley bag label is

$$L_0(\mathbf{w}, \{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}(y_i^0 - \mathbf{w}^T\mathbf{B}_i\boldsymbol{\beta}_i)^2. \tag{3}$$

3.2. Incorporating Knowledge

Ordinal relevance Since the intensity evolves smoothly and increases monotonically in segments, the relevance value depends on the difference between the current frame and the peak or valley frame. The closer the frame is to the peak frame, the larger the peak relevance is for the peak bag label. Similarly, the closer the frame is to the valley frame, the larger the valley relevance is for the valley bag label. In a training segment, the peak relevance increases monotonically, while the valley relevance decreases monotonically. Such knowledge can be leveraged to constrain the relevance values of frames in each segment. For $\mathbf{B}_i$, the feasible domain of the peak relevance $\boldsymbol{\alpha}_i$ is

$$\begin{aligned} S^{\alpha}(\boldsymbol{\alpha}_i) =& \{\boldsymbol{\alpha}_i \in \mathbb{R}^{n_i} | \mathbf{e}_i^T \boldsymbol{\alpha}_i = 1, \\ & 0 \leq \boldsymbol{\alpha}_i^1 \leq \boldsymbol{\alpha}_i^2 \leq ... \leq \boldsymbol{\alpha}_i^{n_i}\}, \end{aligned} \tag{4}$$

where $\mathbf{e}_i$ is an n_i dimensional vector with all elements equal to 1. The feasible domain for the valley relevance $\boldsymbol{\beta}_i$ has the similar constraints except for the ordering,

$$\begin{aligned} S^{\beta}(\boldsymbol{\beta}_i) =& \{\boldsymbol{\beta}_i \in \mathbb{R}^{n_i} | \mathbf{e}_i^T \boldsymbol{\beta}_i = 1, \\ & \boldsymbol{\beta}_i^1 \geq \boldsymbol{\beta}_i^2 \geq ... \geq \boldsymbol{\beta}_i^{n_i} \geq 0\}. \end{aligned} \tag{5}$$

Intensity smoothness The changes of facial appearance are caused by the movements of muscles. The contraction and relaxation of muscles are smooth movements, which lead to the smooth changes of facial appearance. Since the AU intensity is defined according to the local facial appearance, the intensity changes smoothly in expression sequences. The difference between the intensities of neighboring frames should be small. Intensity smoothness can be exploited to provide weak supervision for the joint learning of the relevance and the regressor by encouraging the intensity predictions of neighboring frames to be close. The smoothness can be encoded as a regularization term, *i.e.*,

$$\begin{aligned} R_1(\mathbf{w}, \mathcal{D}) &= \sum_{i=1}^{N} \sum_{j,k=1}^{n_i} \mathbf{C}_i^{j,k} (\mathbf{w}^T \mathbf{B}_i^j - \mathbf{w}^T \mathbf{B}_i^k)^2 \\ &= \frac{1}{2} \mathbf{w}^T \big[\sum_{i=1}^{N} \mathbf{B}_i (\mathbf{D}_i - \mathbf{C}_i) \mathbf{B}_i^T \big] \mathbf{w} = \frac{1}{2} \mathbf{w}^T \mathbf{L} \mathbf{w}, \end{aligned} \tag{6}$$

where $\mathbf{L} = \sum_{i=1}^{N} \mathbf{B}_i (\mathbf{D}_i - \mathbf{C}_i) \mathbf{B}_i^T$. $\mathbf{C}_i$ is the adjacent matrix, where $\mathbf{C}_i^{j,k} = 1$ if $|j-k| = 1$. Otherwise, $\mathbf{C}_i^{j,k} = 0$. $\mathbf{D}_i^{j,j} = \sum_k \mathbf{C}_i^{j,k}$ and $\mathbf{D}_i^{j,k} = 0$ if $j \neq k$.

Relevance smoothness Similar to intensity smoothness, the relevance changes smoothly. The difference between the relevance of neighboring frames should be small. We exploit the relevance smoothness for the peak bag label by using the regularization term, *i.e.*,

$$\begin{aligned} R_2(\{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D}) &= \sum_{i=1}^{N} \sum_{j,k=1}^{n_i} \mathbf{C}_i^{j,k} (\boldsymbol{\alpha}_i^j - \boldsymbol{\alpha}_i^k)^2 \\ &= \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{\alpha}_i^T (\mathbf{D}_i - \mathbf{C}_i) \boldsymbol{\alpha}_i. \end{aligned} \tag{7}$$

The regularization term for the valley bag label is

$$R_2(\{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D}) = \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{\beta}_i^T (\mathbf{D}_i - \mathbf{C}_i) \boldsymbol{\beta}_i. \tag{8}$$

3.3. Complete Formulation

By incorporating the knowledge with weak annotation as described in previous sections, we can jointly learn the relevance and regressor by solving the following problem

$$\begin{aligned} \min_{\mathbf{w}, \{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\}_{i=1}^N} \quad & L(\mathbf{w}, \{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D}) + \lambda_0 L_0(\mathbf{w}, \{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D}) \\ & + \lambda_1 R_1(\mathbf{w}, \mathcal{D}) + \lambda_2 R_2(\{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D}) \\ & + \lambda_3 R_2(\{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D}) + \frac{\lambda_4}{2} \|\mathbf{w}\|^2 \\ s.t. \quad & \boldsymbol{\alpha}_i \in S^{\alpha}(\boldsymbol{\alpha}_i), \boldsymbol{\beta}_i \in S^{\beta}(\boldsymbol{\beta}_i), \\ & \mathbf{V}_i(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i) = \mathbf{0}, i = 1, 2, ..., N, \end{aligned} \tag{9}$$

where $\lambda_k \geq 0, k = 0, 1, 2, 3, 4$, are the penalty parameters. The first term is the loss for the peak bag label. The second term is the loss for the valley bag label. The third term represents the intensity smoothness. The fourth and the fifth represent relevance smoothness. The last term is the regularization on the parameters of the estimator. The constraints represent ordinal relevance.

To convert the ordinal constraints to a compact form for optimization, let $\boldsymbol{\eta}_i = \{\boldsymbol{\eta}_i^1, \boldsymbol{\eta}_i^2, ..., \boldsymbol{\eta}_i^{n_i}\} \in \mathbb{R}^{n_i}$ be the relevance increments in a segment and $\boldsymbol{\eta}_i \geq \mathbf{0}$. The peak relevance can be represented by $\boldsymbol{\alpha}_i = \mathbf{A}_i \boldsymbol{\eta}_i$. $\mathbf{A}_i$ is square matrix with $\mathbf{A}_i^{j,k} = 1$ when $j \geq k$, and $\mathbf{A}_i^{j,k} = 0$ when $j < k$. Similarly, let $\boldsymbol{\mu}_i = \{\boldsymbol{\mu}_i^1, \boldsymbol{\mu}_i^2, ..., \boldsymbol{\mu}_i^{n_i}\} \in \mathbb{R}^{n_i}$ be the relevance increments for the valley bag label and $\boldsymbol{\mu}_i \geq \mathbf{0}$. The valley relevance can be represented by $\boldsymbol{\beta}_i = \mathbf{A}_i^T \boldsymbol{\mu}_i$. Then, the equivalent problem is

$$\begin{aligned} \min_{\mathbf{w}, \{\boldsymbol{\eta}_i, \boldsymbol{\mu}_i\}_{i=1}^N} \quad & L(\mathbf{w}, \{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D}) + \lambda_0 L_0(\mathbf{w}, \{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D}) \\ & + \lambda_1 R_1(\mathbf{w}, \mathcal{D}) + \lambda_2 R_2(\{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D}) \\ & + \lambda_3 R_2(\{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D}) + \frac{\lambda_4}{2} \|\mathbf{w}\|^2 \\ s.t. \quad & \boldsymbol{\eta}_i \geq \mathbf{0}, \boldsymbol{\mu}_i \geq \mathbf{0} \\ & \mathbf{e}_i^T (\mathbf{A}_i \boldsymbol{\eta}_i) = 1, \mathbf{e}_i^T (\mathbf{A}_i^T \boldsymbol{\mu}_i) = 1, \\ & \mathbf{V}_i (\mathbf{A}_i \boldsymbol{\eta}_i + \mathbf{A}_i^T \boldsymbol{\mu}_i) = \mathbf{0} \\ & i = 1, 2, ..., N, \end{aligned} \tag{10}$$

where $\boldsymbol{\alpha}_i = \mathbf{A}_i\boldsymbol{\eta}_i$ and $\boldsymbol{\beta}_i = \mathbf{A}_i^T\boldsymbol{\mu}_i$.

After model learning, we obtain the parameters of the frame-level regressor, i.e., $\mathbf{w}$. We use $y = f(\mathbf{x};\mathbf{w})$ to predict the intensity for an unlabeled frame.

3.4. Optimization

Let $\boldsymbol{\theta}_i = [\boldsymbol{\eta}_i; \boldsymbol{\mu}_i] \in \mathbb{R}^{2n_i}$ by concatenating $\boldsymbol{\eta}_i$ and $\boldsymbol{\mu}_i$. The problem (10) is a biconvex problem in $\mathbf{w}$ and $\boldsymbol{\theta}_i$, since $L(\mathbf{w}, \{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D})$ and $L_0(\mathbf{w}, \{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D})$ are biconvex in $\mathbf{w}$ and $\boldsymbol{\theta}_i$, $R_2(\{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D})$ and $R_2(\{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D})$ are convex in $\boldsymbol{\theta}_i$, and $\|\mathbf{w}\|^2$ is convex in $\mathbf{w}$. To solve problem (10), we develop an iterative optimization algorithm (see Algo. 1) under the alternating minimization framework [5]. Details of derivation are presented in the supplementary material.

Optimize $\mathbf{w}$, given $\{\boldsymbol{\theta}_i\}_{i=1}^N$ Given $\{\boldsymbol{\theta}_i\}_{i=1}^N$, the subproblem with respect to $\mathbf{w}$ becomes

$$\min_{\mathbf{w}} \quad L(\mathbf{w}, \{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D}) + \lambda_0 L_0(\mathbf{w}, \{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D}) + \lambda_1 R_1(\mathbf{w}, \mathcal{D}) + \frac{\lambda_4}{2}\|\mathbf{w}\|^2, \tag{11}$$

where $\{\boldsymbol{\alpha}_i\}_{i=1}^N$ and $\{\boldsymbol{\beta}\}_{i=1}^N$ are known by $\boldsymbol{\alpha}_i = \mathbf{A}_i\boldsymbol{\eta}_i$ and $\boldsymbol{\beta}_i = \mathbf{A}_i^T\boldsymbol{\mu}_i$. The subproblem is an unconstrained quadratic programming problem with respect to $\mathbf{w}$. The solution is

$$\mathbf{w}^* = [\mathbf{X}\mathbf{X}^T + \lambda_0\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \lambda_1\mathbf{L} + \lambda_4\mathbf{I}]^{-1}(\mathbf{X}\mathbf{Y} + \lambda_0\tilde{\mathbf{X}}\mathbf{Y}_0),$$

where $\mathbf{X} = [\mathbf{B}_1\boldsymbol{\alpha}_1, ..., \mathbf{B}_N\boldsymbol{\alpha}_N]$, $\tilde{\mathbf{X}} = [\mathbf{B}_1\boldsymbol{\beta}_1, ..., \mathbf{B}_N\boldsymbol{\beta}_N]$. $\mathbf{Y} = [y_1, ..., y_N]^T$ and $\mathbf{Y}_0 = [y_1^0, ..., y_N^0]^T$ are the peak and valley bag label vectors of the N training segments.

Optimize $\{\boldsymbol{\theta}_i\}_{i=1}^N$, given $\mathbf{w}$ Given $\mathbf{w}$, the subproblem with respect to $\boldsymbol{\theta}_i$ becomes

$$\min_{\boldsymbol{\theta}_i} \quad L(\mathbf{w}, \boldsymbol{\alpha}_i, \mathcal{D}) + \lambda_0 L_0(\mathbf{w}, \boldsymbol{\beta}_i, \mathcal{D}) + \lambda_2 R_2(\boldsymbol{\alpha}_i, \mathcal{D}) + \lambda_3 R_2(\boldsymbol{\beta}_i, \mathcal{D}) \quad s.t. \quad \boldsymbol{\theta}_i \in S^{\theta}(\boldsymbol{\theta}_i), \tag{12}$$

where $S^{\theta}(\boldsymbol{\theta}_i)$ is the set of constraints on $\boldsymbol{\eta}_i, \boldsymbol{\mu}_i$. The problem is a quadratic programming problem with respect to $\boldsymbol{\theta}_i$. Problem (12) can be solved efficiently using quadratic programming algorithm such as [46].

The original problem (10) is decomposed into two subproblems. Both of them are standard quadratic programming which can be solved efficiently. The proposed alternating optimization algorithm is guaranteed to converge since the objective function is minimized at each step and the objective is non-increasing. When optimizing $\{\boldsymbol{\theta}_i\}_{i=1}^N$, the optimization can be perform parallelly for each segment since $\boldsymbol{\theta}_i$ is independent of others given $\mathbf{w}$. The complexity is $\mathcal{O}(nd^2 + d^3 + \sqrt{d}\log(d/\epsilon))$. n, d and ϵ are the number of samples, the feature dimension, and the optimality tolerance. Figure 4 shows the convergence of our method and the learned relevance for two training segments.

Algorithm 1: BORMIR

Input : Training data $\mathcal{D} = \{(\mathbf{B}_i, y_i^0, y_i)\}_{i=1}^N$, $\{\lambda_i\}_{i=0}^4$.

Output: The regressor $\mathbf{w}$, relevance $\{\boldsymbol{\alpha}_i\}_{i=1}^N$, $\{\boldsymbol{\beta}_i\}_{i=1}^N$.

// Assign equal relevance to each frame;

Initialize $\mathbf{w} = \mathbf{0}$, $\boldsymbol{\eta}_i^1 = \frac{1}{n_i}, \boldsymbol{\eta}_i^j = 0, j > 1$; $\boldsymbol{\mu}_i^{n_i} = \frac{1}{n_i}, \boldsymbol{\mu}_i^k = 0, k < n_i$;

while *not converged* **do**

  Fix $\{\boldsymbol{\theta}_i\}_{i=1}^N$, update $\mathbf{w}$ by solving (11) ;

  Fix $\mathbf{w}$, update $\{\boldsymbol{\theta}_i\}_{i=1}^N$ by solving (12) for each training segment;

end

$\boldsymbol{\alpha}_i = \mathbf{A}_i\boldsymbol{\eta}_i, \boldsymbol{\beta}_i = \mathbf{A}_i^T\boldsymbol{\mu}_i, i = 1, ..., N$;

Return $\mathbf{w}$, $\{\boldsymbol{\alpha}_i\}_{i=1}^N$, $\{\boldsymbol{\beta}_i\}_{i=1}^N$

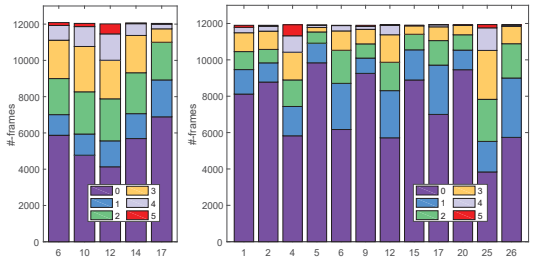

Figure 3. Intensity distribution. Left: FERA 2015. Right: DISFA

4. Experiments

4.1. Settings

Datasets. The proposed method is evaluated on three benchmark databases, *i.e.*, the subset of the Binghamton-Pittsburgh 4D (**FERA 2015**) [39] and **DISFA** [21]. FERA 2015 contains 328 videos from 41 subjects when the subjects are performing 8 tasks. DISFA contains 27 videos from 27 subjects when they are watching YouTube videos. AU intensity is quantified into 6 discrete levels. In FERA 2015, we use the official Training/Development splits for evaluation. In DISFA, we perform 5-fold subject independent cross-validation. To compare with MI-DORF [30], we also perform an experiment on the UNBC-MacMaster Shoulder Pain Expression Archive (**PAIN**) [18] for pain intensity estimation. The PAIN database contains videos of 25 patients when they move their shoulders. Pain intensity is quantified into 16 levels, but only few frames have the highest intensity levels. We follow the strategy in [28] to group intensity levels, *i.e.*, 0 (0), 1 (1), 2 (2), 3 (3), 4 (4$\sim$5), 5 (6 $\sim$ 15). In PAIN, we perform leave-one-subject-out cross validation as in [30].

The weak annotation contains the locations and intensity annotations of the peak and valley frames. Obtaining the weak annotation requires much less effort than anno-

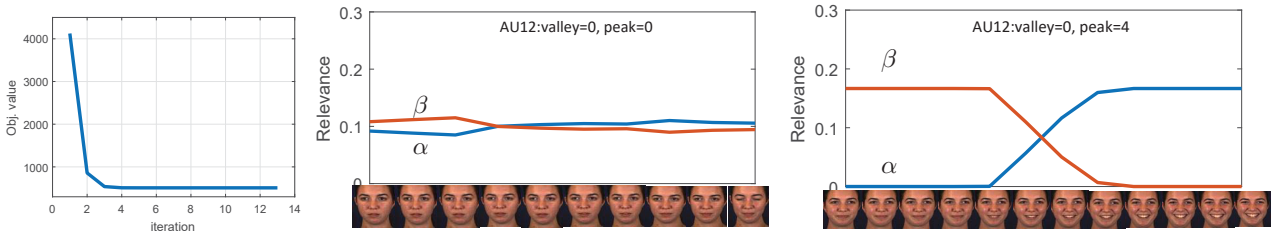

Figure 4. Left: the convergence of BORMIR. Right: the learned relevance of each frame in two segments.

Table 1. Comparison to the baseline methods.

Database		FERA 2015						DISFA												
AU		6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
PCC	BORMIR-DS	.652	.617	.808	.300	.470	.569	.208	.154	**.369**	.216	.284	.061	.394	.118	.070	.077	.552	.097	.217
	BORMIR-DO	.718	.674	.853	.377	**.494**	.623	.246	**.351**	.323	.220	.430	.255	.633	**.289**	.314	.172	.708	.129	.339
	BORMIR-DB	.699	.681	.854	.371	.485	.618	.242	.345	.330	.220	**.448**	**.277**	**.642**	.283	**.324**	.187	.727	.143	.347
	BORMIR	**.729**	**.689**	**.865**	**.400**	.493	**.635**	**.259**	.333	.360	**.251**	.447	.244	.629	.287	.316	**.207**	**.733**	**.168**	**.353**
ICC	BORMIR-DS	.635	.616	.801	.299	.469	.564	.176	.092	.296	.150	.249	.038	.359	.083	.063	.054	.498	.086	.179
	BORMIR-DO	.718	.666	.852	.358	.476	.614	.165	.245	.266	.152	.355	.168	.551	**.158**	.216	.074	.667	.110	.261
	BORMIR-DB	.696	**.681**	.852	.366	**.483**	.616	.156	.232	.268	.142	.365	.164	.571	.146	.219	.074	.690	.122	.262
	BORMIR	**.725**	.675	**.861**	**.368**	.469	**.620**	**.198**	**.248**	**.302**	**.173**	**.385**	**.181**	**.583**	.157	**.225**	**.088**	**.707**	**.148**	**.283**
MAE	BORMIR-DS	1.085	1.082	.951	1.342	.935	1.079	1.703	2.374	2.659	1.564	1.484	2.512	1.610	1.609	1.510	1.460	1.784	1.576	1.820
	BORMIR-DO	.876	.919	.756	1.076	.809	.887	1.016	.966	1.275	.820	.795	1.029	.785	.699	.791	.831	.877	.925	.901
	BORMIR-DB	1.198	1.075	.776	1.212	.967	1.046	1.084	1.016	1.347	.850	.816	1.032	.825	.735	.830	.849	**.845**	.958	.932
	BORMIR	**.848**	**.895**	**.678**	**1.046**	**.791**	**.852**	**.875**	**.783**	**1.240**	**.589**	**.769**	**.777**	**.757**	**.564**	**.716**	**.628**	.898	**.875**	**.789**

tating every frame in sequences. The locations are identified manually according to the definitions of peak and valley frames in [20]. Since sequences are captured at a high frame rate, faces in the consecutive frames do not have distinct changes. The sequences are downsampled. The distributions of AU intensity levels are shown in Figure 3. PAIN has 6497 frames and the distribution of pain intensity is 0 (68.5%), 1 (10%), 2 (8.9%), 3 (5.6%), 4 (4.1%), 5 (2.2%). The average numbers of annotated frames are around 850 for FERA 2015, 900 for DISFA, and 350 for PAIN.

Image features. Both databases provide 66 facial landmarks. [39] provides a way to extract geometric features from 49 inner facial landmarks. Facial shapes are firstly aligned by using several stable points, *i.e.*, points of eye corners and nose. By subtracting the mean shape from the aligned faces, we can get 98D features. By computing distance between two consecutive points and angles among three consecutive points, we can get another 71D features. By computing the distance between a point to the median of the stable points, we can get another 49D features. We then concatenate these features to be 218D features and apply Gaussian normalization to each dimension.

Evaluation metrics and hyperparameters. We use three metrics for evaluation, *i.e.*, Pearson correlation coefficient (PCC), intra-class correlation (ICC(3,1) [34]), and mean absolute error (MAE). PCC is used to measure the linear association between the ground truth and the predicted intensity. ICC is commonly used to measure the agreement between annotators. MAE is commonly used for ordinal prediction tasks [15, 43]. Our model has five hyperparameters $\{\lambda_i\}_{i=0}^4$, which are selected through validation. The training set is split into two parts, *i.e.*, one with 60% sequences as the training set and the other with 40% as the validation set. We use grid search strategy for parameter selection. The ranges are $\lambda_0 \in \{0.2, 0.4, 0.6, 0.8, 1\}$ and $\{\lambda_i\}_{i=1}^4 \in \{10^{-2}, 10^{-1}, 10^0, 10^1\}$. We use MAE as the measure to evaluate the performance on the labeled frames in the validation set for parameter selection.

Comparative methods. We first compare the performance of our method (BORMIR) with its variants which drop one type of knowledge. BORMIR exploits all types of knowledge. BORMIR-DS drops the smoothness of relevance and intensity. BORMIR-DO drops the ordinal relevance. BORMIR-DB drops the valley bag label. Then, we compare our method to the state-of-the-art methods, including supervised methods and weakly supervised methods. SOVR [4], RVR [12], LT [13], and DSRVM [14] are supervised methods. SOVR is used for ordinal regression while others are the state-of-the-art AU intensity estimation methods. These supervised methods use only the annotations of peak and valley frames since they can handle unlabeled frames. MIR [40], OSVR [50], and MI-DORF [30] are weakly supervised methods, which use the annotations of peak and valley frames and also unlabeled frames. Furthermore, we also compare our method to the state-of-the-art supervised deep models such as CNN [9], CCNN-IT [42], and 2DC [17]. For deep models, every frame in sequences is annotated with AU intensity.

4.2. Results

Comparison with the baseline methods. The results are shown in Table 1. Our method achieves better average performance than its three variants under all evaluation metrics on both FERA 2015 and DISFA. Our method incorporates all types of knowledge while each variant drops one type of knowledge. The results show that each type of

Table 2. Comparison to the state-of-the-art methods. The scenario is that the training set is partially annotated. Numbers in bracket and bold represent the best performance; numbers in bold only represent the second best.

Database		FERA 2015						DISFA												
	AU	6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
PCC	SOVR [4]	.613	.550	.781	.293	.443	.536	.211	.090	[**.368**]	.213	.329	.078	.391	**.141**	.084	.141	.702	**.221**	.247
	RVR [12]	**.676**	.614	**.815**	.311	**.485**	**.580**	**.277**	**.131**	**.364**	**.313**	.315	.109	.450	.078	**.114**	[**.231**]	**.716**	.201	**.275**
	LT [13]	.578	**.622**	.710	.334	.116	.472	[**.280**]	.017	.078	.113	.313	.069	.524	.100	.038	-.034	.342	[**.238**]	.173
	DSRVM [14]	.604	.585	.778	**.343**	.341	.530	.131	-.016	.083	.067	**.391**	.115	**.607**	-.013	-.084	.024	.646	.059	.168
	MIR [40]	.567	.466	.656	.299	.314	.460	.196	.114	.312	[**.318**]	.258	.188	.426	.078	.100	.085	.622	.072	.231
	OSVR [50]	.647	.577	.783	.271	.449	.545	.238	.074	.332	.215	.264	**.199**	.342	.130	.079	.114	.664	.117	.231
	BORMIR	[**.729**]	[**.689**]	[**.865**]	[**.400**]	[**.493**]	[**.635**]	.259	[**.333**]	.360	.251	[**.447**]	[**.244**]	[**.629**]	[**.287**]	[**.316**]	**.207**	[**.733**]	.168	[**.353**]
ICC	SOVR [4]	.610	.549	.775	.293	.441	.534	.195	.048	**.293**	.131	**.297**	.039	.363	**.123**	.077	**.114**	.660	[**.197**]	.211
	RVR [12]	**.675**	**.609**	**.814**	**.311**	[**.481**]	**.578**	[**.244**]	**.089**	.288	[**.203**]	.286	.067	.432	.076	**.109**	[**.216**]	**.679**	**.180**	**.239**
	LT [13]	.558	.587	.695	.292	.094	.445	**.216**	.017	.035	.102	.230	.043	.434	.043	.021	-.025	.289	.140	.129
	DSRVM [14]	.600	.569	.772	.290	.299	.506	.043	-.012	.026	.039	.295	.063	**.553**	.000	-.022	.002	.632	.056	.140
	MIR [40]	.438	.394	.600	.229	.230	.378	.117	.062	.199	**.186**	.200	.124	.328	.051	.063	.049	.477	.046	.159
	OSVR [50]	.646	.577	.780	.269	.449	.544	.208	.038	.248	.151	.229	**.152**	.313	.115	.066	.094	.618	.093	.194
	BORMIR	[**.725**]	[**.675**]	[**.861**]	[**.368**]	**.469**	[**.620**]	.198	[**.248**]	[**.302**]	.173	[**.385**]	[**.181**]	[**.583**]	[**.157**]	[**.225**]	.088	[**.707**]	.148	[**.283**]
MAE	SOVR [4]	1.080	1.176	.967	1.314	.902	1.088	1.612	2.153	2.661	1.768	1.311	2.024	1.481	1.442	1.430	1.512	1.182	1.455	1.669
	RVR [12]	.959	1.030	**.838**	1.218	**.837**	.976	1.633	1.781	2.574	1.534	1.357	2.077	1.285	1.237	1.343	.967	1.234	1.426	1.537
	LT [13]	.948	**.978**	.911	1.116	.958	.982	1.067	.958	**1.370**	**.544**	**.806**	.902	**.882**	.606	.771	.668	1.241	[**.802**]	.885
	DSRVM [14]	**.944**	1.017	**.838**	**1.086**	.856	**.948**	[**.821**]	[**.619**]	1.453	[**.332**]	.891	[**.553**]	.905	[**.441**]	[**.690**]	[**.371**]	**1.071**	.943	[**.757**]
	MIR [40]	1.992	2.072	1.502	2.315	1.957	1.968	2.682	2.668	3.905	1.888	1.957	2.173	2.062	1.965	2.488	1.964	2.191	2.135	2.340
	OSVR [50]	1.024	1.126	.953	1.354	.928	1.077	1.648	1.873	2.943	1.378	1.556	1.690	1.636	1.101	1.614	1.371	1.329	1.789	1.661
	BORMIR	[**.848**]	[**.895**]	[**.678**]	[**1.046**]	[**.791**]	[**.852**]	**.875**	**.783**	[**1.240**]	.589	[**.769**]	**.777**	[**.757**]	**.564**	**.716**	**.628**	[**.898**]	**.875**	**.789**

knowledge can help improve the performance. Compared to BORMIR-DS, when dropping the smoothness, the performance decreases significantly. It shows that the temporal smoothness is relatively more important than other two types of knowledge when the annotations are limited.

Comparison with the-state-of-the-art methods. The results are shown in Table 2. Our method outperforms other competing methods on both databases under all evaluation metrics except MAE on DISFA, where ours is the second best. The comparisons are analyzed as follows. Firstly, our method achieves much better performance than MIR. MIR uses only the peak bag label and does not exploit any knowledge. It performs poorly for frame-level prediction. The result further demonstrates the effectiveness of the introduced valley bag label and the domain knowledge. Secondly, among the supervised methods, RVR achieves better results than others. Compared to RVR, our method achieves the improvement of over 5% under PCC and 4% under ICC and also better MAE, especially on DISFA. LT and DSRVM do not perform well on DISFA which has an imbalanced distribution of intensity levels. Our method can exploit both labeled and unlabeled frames through domain knowledge while supervised methods can only use labeled frames. Thirdly, our method also outperforms the weakly supervised methods such as OSVR, especially on DISFA. Compared to OSVR, we formulate AU intensity estimation from a new perspective by introducing the relevance to model sequential data. Besides, we incorporate more types of knowledge to exploit unlabeled frames.

Table 3 shows the results of pain intensity estimation on PAIN. It also contains the reported performance of several deep models and state-of-the-art methods which use the annotations of all frames for training. (*) indicates that the results are adapted from the corresponding paper. Compared to MI-DORF, it outputs discrete intensity values while ours is continuous. Although MAE of MI-DORF is smaller than ours, we can better capture the trend with higher PCC and ICC. Compared to the methods that using the fully annotated database, our method achieves promising results in PCC and MAE when less than 10% of frames are annotated.

Table 3. Performance of different methods for pain intensity estimation on the PAIN database.

Method	PCC	ICC	MAE
Fully annotated			
RCNN [51]*	.650	-	-
FCNN [7]*	.673	-	-
KCORFh [28]*	-	.703	.800
csCORFwh [29]*	-	.640	.820
Partially annotated			
MI-DORF [30]*	.460	.460	**.510**
BORMIR	**.605**	**.531**	.821

Table 4. Comparison to fully supervised deep models.

Data	FERA 2015			DISFA		
Method	PCC	ICC	MAE	PCC	ICC	MAE
Fully annotated						
CCNN-IT [42]*	–	.630	1.260	–	.380	.660
2DC [17]*	–	**.660**	–	–	**.500**	–
CNN [9]	**.638**	.632	**.783**	.324	.305	**.496**
Partially annotated						
BORMIR	.635	.620	.852	**.353**	.283	.789

Results of comparison with the state-of-the-art deep models are shown in Table 4. The average performance is reported. Though our method use substantially less amount of annotations, it achieves comparable performance to the deep models on FERA 2015. It also achieves promising performance on DISFA. Deep models require plenty of annotations for training to avoid overfitting while our method needs only the annotations of key frames which occupy a small portion of frames in databases.

Performance of using partial training segments. To further evaluate the effectiveness of the proposed method,

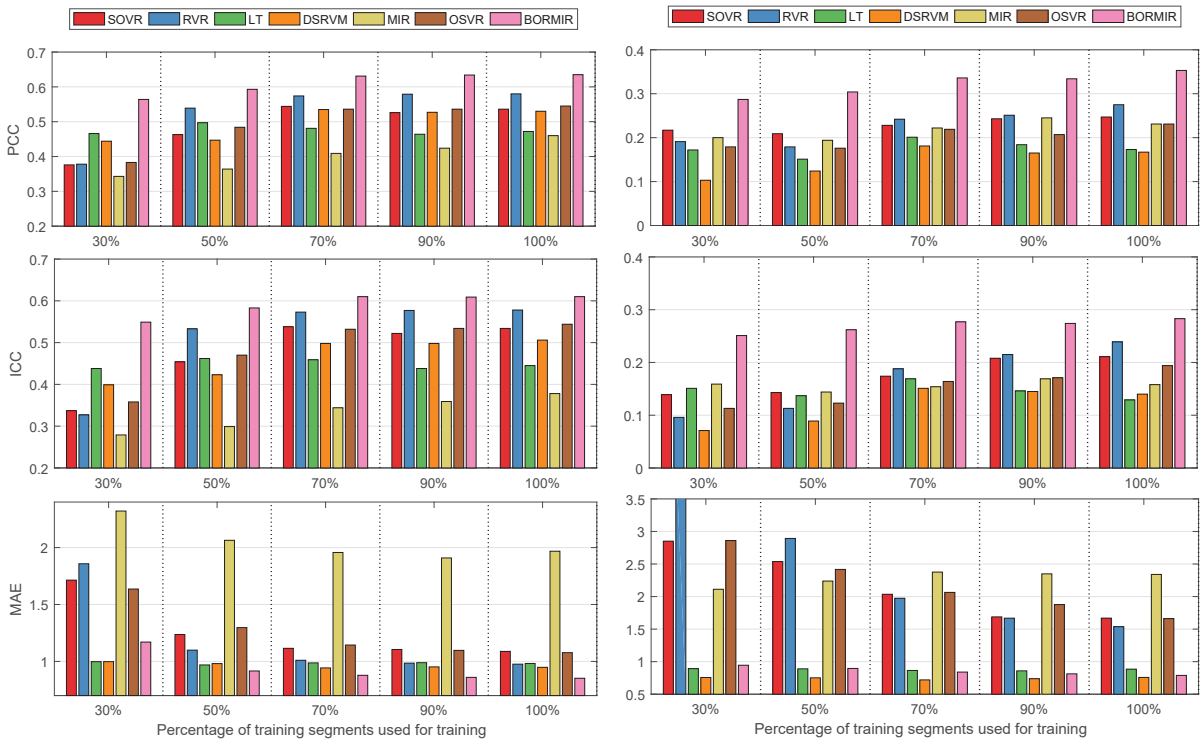

Figure 5. Using partial training segments. Left: FERA 2015. Right: DISFA. The results are the average performance of all the AUs.

we reduce the number of training segments to evaluate the performance of different methods. We consider the scenarios that 30%, 50%, 70%, 90%, and 100% of training segments are used for training. Note that only the peak and valley frames in each segment are annotated. The results are shown in Figure 5. On both FERA 2015 and DISFA, our method outperforms competing methods in PCC and ICC under all the scenarios. On FERA 2015, our method also outperforms other methods in MAE under the scenarios expect 30% case. On DISFA, the MAE of our method is better than other methods expect for DSRVM. However, DSRVM has poor performance in PCC and ICC. As shown in Figure 5, when reducing the number of training segments, the performances of competing methods decrease substantially while our method can still perform well.

Robustness to noisy annotations. Our method does not explicitly consider spurious peaks and valleys. To evaluate its robustness, we add noise to the annotations of locations of peaks and valleys by shifting their original location annotations. We perform an experiment on FERA 2015. The experiment setting is the same as AU intensity estimation except for shifted original location annotations. When the shift is 10 frames, the average performance is (PCC: 0.629, ICC: 0.605, MAE: 0.873). When the shift is 30 frames, it becomes (PCC: 0.539, ICC: 0.510, MAE: 0.937). When the shift is large, the weak supervision will provide incorrect information. The results show that our method can tolerant noisy annotations to some extent.

5. Conclusion

We propose a novel weakly supervised learning approach, BORMIR, which can learn frame-level intensity regressor with weakly labeled sequences. The weak annotation is much easier to obtain than annotating the intensity of every frame. We study AU intensity estimation from a new perspective by introducing the concept of 'relevance' to model sequential data. Besides, we simultaneously consider two bag labels for one bag and each frame has two relevance values associated with the two bag labels. The AU intensity is formulated as a joint learning problem of the relevance and intensity regressor. To make the learning of frame-level intensity regressor feasible, we incorporate domain knowledge on the relevance and AU intensity to provide weak supervision and exploit unlabeled samples. We also propose an efficient algorithm for optimization. Evaluations on three benchmark databases demonstrate the effectiveness of the proposed method.

Acknowledgments: The work was accomplished when the first author was visiting Rensselaer Polytechnic Institute (RPI), through a scholarship from China Scholarship Council. The support of CSC and RPI is greatly appreciated. This work was also supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61672520, 61573348 and 61702488.

References

[1] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Continuous conditional neural fields for structured regression. In *ECCV*, 2014. 2

[2] J. Chen, S. Nie, and Q. Ji. Data-free prior model for upper body pose estimation and tracking. *TIP*, 2013. 2

[3] P.-M. Cheung and J. T. Kwok. A regularization framework for multiple-instance learning. In *ICML*, 2006. 3

[4] W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. In *ICML*, 2005. 6, 7

[5] I. Csisz, G. Tusnády, et al. Information geometry and alternating minimization procedures. *Statistics and decisions*, 1984. 5

[6] F. De la Torre, T. Simon, Z. Ambadar, and J. F. Cohn. Fast-facs: A computer-assisted system to increase speed and reliability of manual facs coding. In *ACII*, 2011. 2

[7] J. Egede, M. Valstar, and B. Martinez. Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. *arXiv preprint arXiv:1701.04540*, 2017. 2, 7

[8] P. Ekman and W. V. Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978. 1

[9] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *FG workshop*, volume 6, 2015. 1, 2, 6, 7

[10] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De La Torre. Continuous au intensity estimation using localized, sparse facial feature space. In *FG*, 2013. 2

[11] L. A. Jeni, A. Lőrincz, T. Nagy, Z. Palotai, J. Sebők, Z. Szabó, and D. Takács. 3d shape estimation in video sequences provides high precision evaluation of facial expressions. *IVC*, 2012. 2

[12] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *ISVC*, 2012. 1, 2, 6, 7

[13] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, 2015. 2, 6, 7

[14] S. Kaltwang, S. Todorovic, and M. Pantic. Doubly sparse relevance vector machine for continuous facial behavior estimation. *TPAMI*, 2016. 2, 6, 7

[15] M. Kim and V. Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. In *ECCV*, 2010. 6

[16] Y. Li, S. M. Mavadati, M. H. Mahoor, Y. Zhao, and Q. Ji. Measuring the intensity of spontaneous facial action units with dynamic bayesian network. *PR*, 2015. 2

[17] D. Linh Tran, R. Walecki, O. (Oggi) Rudovic, S. Eleftheriadis, B. Schuller, and M. Pantic. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *ICCV*, 2017. 2, 6, 7

[18] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *FG workshop*, 2011. 5

[19] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *CVPRW*, 2009. 2

[20] M. Mavadati, P. Sanger, and M. H. Mahoor. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In *CVPRW*, 2016. 2, 3, 6

[21] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *TAC*, 2013. 5

[22] Z. Ming, A. Bugeau, J.-L. Rouas, and T. Shochi. Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine. In *FG Workshops*, 2015. 2

[23] J. Nicolle, K. Bailly, and M. Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *FG Workshop*, 2015. 2

[24] S. Nie and Q. Ji. Capturing global and local dynamics for human action recognition. In *ICPR*, 2014. 2

[25] S. Nie, Z. Wang, and Q. Ji. A generative restricted boltzmann machine based method for high-dimensional motion data modeling. *CVIU*, 136, 2015. 2

[26] S. Ray and D. Page. Multiple instance regression. In *ICML*, volume 1, pages 425–432, 2001. 3

[27] O. Rudovic, V. Pavlovic, and M. Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *CVPR*, 2012. 2

[28] O. Rudovic, V. Pavlovic, and M. Pantic. Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields. In *ISVC*, 2013. 2, 5, 7

[29] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *TPAMI*, 2015. 2, 7

[30] A. Ruiz, O. Rudovic, X. Binefa, and M. Pantic. Multi-instance dynamic ordinal random fields for weakly-supervised facial behavior analysis. *arXiv preprint arXiv:1803.00907*, 2018. 1, 2, 3, 5, 6, 7

[31] A. Ruiz, J. Van de Weijer, and X. Binefa. Regularized multi-concept mil for weakly-supervised facial behavior categorization. In *BMVC*, 2014. 2

[32] G. Sandbach, S. Zafeiriou, and M. Pantic. Markov random field structures for facial action unit intensity estimation. In *ICCVW*, 2013. 2

[33] A. Savran, B. Sankur, and M. T. Bilge. Regression-based intensity estimation of facial action units. *IVC*, 2012. 2

[34] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 1979. 6

[35] K. Sikka, A. Dhall, and M. Bartlett. Weakly supervised pain localization using multiple instance learning. In *FG workshop*, 2013. 2

[36] K. Sikka, G. Sharma, and M. Bartlett. Lomo: Latent ordinal model for facial analysis in videos. In *CVPR*, 2016. 2

[37] D. M. Tax, E. Hendriks, M. F. Valstar, and M. Pantic. The detection of concept frames using clustering multi-instance learning. In *ICPR*, 2010. 2

[38] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *TPAMI*, 2007. 2

[39] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *FG workshop*, 2015. 2, 5, 6

[40] K. L. Wagstaff and T. Lane. *Salience assignment for multiple-instance regression*. Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration, 2007. 3, 6, 7

[41] K. L. Wagstaff, T. Lane, and A. Roper. Multiple-instance regression with structured data. In *ICMD workshop*, 2008. 3

[42] R. Walecki, V. Pavlovic, B. Schuller, M. Pantic, et al. Deep structured learning for facial action unit intensity estimation. *arXiv preprint arXiv:1704.04481*, 2017. 2, 6, 7

[43] R. Walecki, O. Rudovic, M. Pantic, and V. Pavlovic. Copula ordinal regression for joint estimation of facial action unit intensity. In *CVPR*, 2016. 2, 6

[44] S. Wang, J. Yang, Z. Gao, and Q. Ji. Feature and label relation modeling for multiple-facial action unit classification and intensity estimation. *Pattern Recognition*, 2017. 2

[45] Z. Wang, V. Radosavljevic, B. Han, Z. Obradovic, and S. Vucetic. Aerosol optical depth prediction from satellite observations by multiple instance regression. In *ICDM*, 2008. 3

[46] S. J. Wright. *Primal-dual interior-point methods*. SIAM, 1997. 5

[47] B. Wu, B.-G. Hu, and Q. Ji. A coupled hidden markov random field model for simultaneous face clustering and tracking in videos. *Pattern Recognition*, 2017. 2

[48] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Simultaneous clustering and tracklet linking for multi-face tracking in videos. In *ICCV*, 2013. 2

[49] B. Wu, Y. Zhang, B.-G. Hu, and Q. Ji. Constrained clustering and its application to face clustering in videos. In *CVPR*, 2013. 2

[50] R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial expression intensity estimation using ordinal information. In *CVPR*, 2016. 1, 2, 3, 6, 7

[51] J. Zhou, X. Hong, F. Su, and G. Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *CVPRW*, 2016. 7

[52] Y. Zhu, F. De la Torre, J. F. Cohn, and Y.-J. Zhang. Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection. In *ACII*, 2009. 2