Multi-Instance Dynamic Ordinal Random Fields for Weakly Supervised Facial Behavior Analysis

Adria Ruiz, Ognjen Rudovic, Xavier Binefa, and Maja Pantic

***Abstract*— We propose a multi-instance-learning (MIL) approach for weakly supervised learning problems, where a training set is formed by bags (sets of feature vectors or instances) and only labels at bag-level are provided. Specifically, we consider the multi-instance dynamic-ordinal-regression (MI-DOR) setting, where the instance labels are naturally represented as ordinal variables and bags are structured as temporal sequences. To this end, we propose MI dynamic ordinal random fields (MI-DORF). In this paper, we treat instance-labels as temporally dependent latent variables in an undirected graphical model. Different MIL assumptions are modelled via newly introduced high-order potentials relating bag and instance-labels within the energy function of the model. We also extend our framework to address the partially observed MI-DOR problem, where a subset of instance labels is also available during training. We show on the tasks of weakly supervised facial action unit and pain intensity estimation, that the proposed framework outperforms alternative learning approaches. Furthermore, we show that MI-DORF can be employed to reduce the data annotation efforts in this context by large-scale.**

***Index Terms*— Mutiple instance learning, undirected graphical models, facial behavior analysis, pain intensity, action units.**

I. Introduction

MUTLI-INSTANCE-LEARNING (MIL) is a popular modelling framework for addressing different weakly-supervised problems [1]–[3]. In traditional Single-Instance-Learning (SIL), the fully supervised setting is assumed with the goal to learn a model from a set of feature vectors (instances) each being annotated in terms of target label y. By contrast, in MIL, the weak supervision is assumed, thus, the training set is formed by bags (sets of instances), and only labels at bag-level are provided. In order to learn a model from this weak-information, MIL assumes that there exists an underlying relation between the label of a bag (e.g., video) and the labels of its constituent instances (e.g., image frames). For instance, in standard Multi-Instance-Classification (MIC) [4], labels are considered binary variables $y \in \{-1, 1\}$ and negative bags are assumed to contain only instances with an associated negative label. In contrast, positive bags must contain at least one positive instance. Another example of MIL assumption is related to the Multi-Instance-Regression (MIR) problem [5], where $y \in R$ is a real-valued variable and the maximum instance-label within the bag is assumed to be equal to y. Different from previous works, in this paper we focus on a novel MIL problem that we refer to as Multi-Instance Dynamic Ordinal Regression (MI-DOR). In this case, bags are structured as dynamic sequences of instances with temporal dependencies. Moreover, instance labels are considered ordinal variables which can take values in a set of L discrete categories satisfying the increasing monotonicity constraints $\{0 \prec \ldots \prec l \prec L\}$. Our definition of MI-DOR is enough general to define different weak-relations between bag and instance-labels. Specifically, we focus on two instances of this problem: Maximum and Relative MI-DOR. Similar to MIR, in the former, we assume that the maximum ordinal value within a sequence is equal to its bag (sequence) label. On the other hand, the latter assumes that the weak-label provides information about the evolution (increase, decrease or monotone) of the instance ordinal levels within the sequence. As we discuss below, these two have important applications in the context of Facial Behavior Analysis that we address in this paper.

A. Motivation: Weakly-Supervised Facial Behavior Analysis

Facial expressions provide information about human emotions, attitudes and mental states [6]. Their automatic analysis has become a very active research field in Computer Vision in the last decade due to the large number of potential applications in different contexts such as medicine or entertainment. In this work, we focus on two relevant problems of automatic facial behavior analysis: Action Unit (AU) [7] and Pain [8] Intensity estimation. Both can be naturally posed as Dynamical Ordinal Regression problems, where the goal is to predict a value on an ordinal scale for each instant of a sequence. Specifically, in AU intensity estimation, the objective is to predict the activation level (on a six-point ordinal scale) of

Manuscript received July 6, 2017; accepted April 13, 2018. Date of publication April 25, 2018; date of current version May 16, 2018. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant 645012. The work of O. Rudovic was supported by the H2020 Research Program through the Marie Skodowska-Curie Grant Agreement under Grant 701236 (EngageME). The work of X. Binefa was supported in part by Spanish Government under Grant MINECO TIN2017-90124-P and in part by the Generalitat de Catalunya under Grant MINECO TIN2017-90124-P and Grant AGAUR 2017-SGR-1311. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ling Shao. *(Corresponding author: Adria Ruiz.)*

A. Ruiz was with the Cognitive Media Technology Group, Universitat Pompeu Fabra, 08018 Barcelona, Spain. He is now with the THOTH Group, INRIA, 38330 Grenoble, France (e-mail: adria.ruiz-ovejero@inria.fr).

O. Rudovic is with the MIT Media Laboratory, Cambridge, MA 02139 USA (e-mail: orudovic@mit.edu).

X. Binefa is with the Cognitive Media Technology Group, Universitat Pompeu Fabra, 08018 Barcelona, Spain (e-mail: xavier.binefa@upf.edu).

M. Pantic is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K., and also with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7522 NB Enschede, The Netherlands (e-mail: m.pantic@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2830189

1057-7149 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

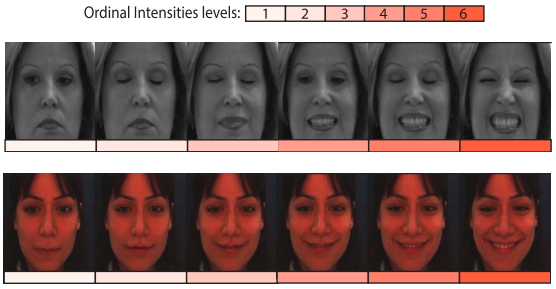

Fig. 1. Illustration of the Pain and Action Unit intensity problems addressed in this work. Top: Sequence showing different pain levels (coded in an ordinal scale from 1 to 6). Bottom: Example of different intensities for Action Unit 12 (Lip-Corner Puller) also represented in an ordinal scale.

facial actions at each frame in a video. Similarly, in the Pain Intensity estimation task we aim to measure the intensity level of pain felt by a patient (see Fig. 1).

The majority of proposed approaches to these problems have followed the supervised learning paradigm [9]–[11], i.e, models are learned using manually annotated labels for each frame in a set of training sequences. Despite the efforts in the field, performance of current approaches following this strategy can still be considered far from optimal. Perhaps, the main reason for such low performance is the limited data used to train supervised models. Annotation in facial behavior analysis is usually a time-consuming task. For example, labeling AU activation levels in one minute of video can require one hour for a specially trained coder. As a consequence, current datasets are sub-optimal in terms of size/variability and, therefore, the use of this limited data for training supervised models can decrease their performance in unseen test samples. Previous works such as [12] have provided empirical evidences supporting this hypothesis.

One potential solution addressing this limitation could be to annotate larger training sets. However, this strategy is not feasible given the expense of the annotation process. In contrast, the explored solution in this work consists of using the weakly-supervised paradigm instead of the fully-supervised one. Weakly-supervised approaches aim to learn models using annotations which only provide partial information (weak-labels) about the task that needs to be solved. These weak-labels are much easier to obtain than those for fully-supervised learning, thus allowing us to use larger datasets minimizing the annotation effort. For example, in Pain Intensity estimation, it is much easier to obtain a label for the whole sequence in terms of the maximum pain intensity felt by the recorded subject (e.g. using patients self-reports or external observers). Similarly, annotating Facial Action Unit intensities requires a huge effort by expert coders. In contrast, segmenting sequences according to the increasing or decreasing evolution of AU intensities (i.e, onset and appex segments) is less time-consuming. These two scenarios motivates our interest in the Maximum and Relative MI-DOR problems previously introduced. Models able to learn only from these weak information would allow to leverage larger training sets and thus potentially build more effective models for intensity estimation of different facial behaviours.

B. Contributions

In this work, we propose the Multi-Instance Dynamic Ordinal Random Fields (MI-DORF) framework to address MI-DOR problems. To build our approach, we use the notion of Hidden Conditional Ordinal Random Fields (HCORF) [13]. Similar to HCORF, MI-DORF is an Undirected Graphical Model where observation labels are modelled as a linear-chain of ordinal latent variables. However, the energy function of MI-DORF is designed to explicitly incorporate the Multiple Instance relation between latent instance labels and observable sequence weak-labels. The main contributions of this work can be summarized as follows:

- To the best of our knowledge, no previous works have explored Multi-Instance Dynamic Ordinal Regression problems (Sec. III). The proposed MI-DORF framework addresses these tasks by explicitly modelling the weak-relation between instances and sequence labels. Our framework is the first MIL approach that imposes ordinal constraints on the instance labels. The proposed method also incorporates dynamic information that is important when modeling temporal structure in instances within the bags (i.e., image sequences). While modeling dynamic information has been attempted in [14] and [15], there are virtually no works that account for both ordinal and temporal data structures within existing MIL frameworks.
- We also introduce high-order potentials in the MI-DORF energy function in order to model weakly-supervised MIL assumptions. Following this strategy, we present two variants of this framework: MaxMI-DORF (Sec. IV) and RelMI-DORF (Sec. V). A preliminary version of the particular MaxMI-DORF method was presented in our previous work [16]. These two models are specially designed to address the Maximum and Relative MI-DOR problems, respectively. Given that the newly introduced MIL potentials of our models render the standard inference procedures for existing latent variable models (e.g., HCORF) infeasible, we derive a novel inference procedure. This procedure scales well with the data number and its computational complexity is similar to that of forward-backward algorithm [17], typically employed in linear-chains models.
- We also propose the Partially-Observed extension of our MI-DORF model (Sec VI). This approach allows us to leverage available instance labels in order to increase the level of supervision in our model. To this end, we generalize the learning and inference procedures of the MI-DORF models mentioned above, making them applicable to the partially-observed and still weakly-supervised learning tasks. We show that with a small portion of labeled instances, we can reach the performance of the fully supervised models for target tasks, thus, reducing the expensive (manual) data annotation efforts by large-scale.

We demonstrate the performance of the proposed methods on weakly-supervised Pain (Sec. VII-E) and Action Unit Intensity estimation using the benchmark datasets for target tasks (Sec. VII-D). We show under various settings the

advantages of our method compared to alternative approaches.

II. Related Work

A. Multiple-Instance Learning

Existing MIL approaches usually follow the bag-based or instance-based paradigms [18]. In the bag-based methods, a feature vector representation for each bag is first extracted. Then, these representations are used to train standard Single-Instance methods, used to estimate the bag labels. This representation is usually computed by using different types of similarity metrics between training instances. Examples following this paradigm include Multi-Instance Kernel [19], MILES [20] or MI-Graph [21]. The main limitation of these approaches is that the learned models can only make predictions at the bag-level (e.g., a video) and are not able to estimate instance-labels (e.g., frame-level intensities). In contrast, instance-based methods directly learn a model which operates at the instance level. For this, MIL assumptions are incorporated by considering instance-labels as latent variables. Using this strategy, traditional supervised models are adapted to incorporate MIL assumptions. Examples of methods following this approach include Multi-Instance Support Vector Machines [22] (MI-SVM), MILBoost [23], MI Gaussian Processes [24] or MI Logistic Regression [25]. In this work, we follow the instance-based paradigm by treating instance-labels as ordinal latent states in a Latent-Dynamic Model. In particular, we follow a similar idea to that in the Multi-Instance Discriminative Markov Networks [26], where the energy function of a Markov Network is designed to explicitly model weak-relations between bag and instance labels. However, in contrast to the works described above, the presented MI-DORF framework accounts for the ordinal structure in instance labels, while also accounting for their dynamics.

B. Latent-Dynamic Models

Popular methods for sequence classification are Latent-Dynamic Models such as Hidden Conditional Random Fields (HCRFs) [27] or Hidden-Markov-Models (HMMs) [28]. These methods are variants of Dynamic Bayesian Networks (DBNs) where a set of latent states are used to model the conditional distribution of observations given the sequence label. In these approaches, dynamic information is modelled by incorporating probabilistic dependence between time-consecutive latent states. MI-DORF builds upon the HCORF framework [13] which considers latent states as ordinal variables. However, HCORF follows the supervised paradigm, where the main goal is to predict sequence labels and latent variables are only used to increase the expressive power of the model. In contrast, the energy function of MI-DORF is defined to explicitly encode Multi-Instance relationships between bag and latent instance labels. Note also that more recent works (e.g., [14], [15]) extended HMMs/HCRFs, respectively, for Multi Instance Classification. The reported results in these works suggested that modeling dynamics in MIL can be beneficial when bag-instances exhibit temporal structure. However, these methods limit their consideration to the case where instance labels are binary and, therefore, are unable to solve MI-DOR problems.

As has been introduced in Sec. I-B, we also extend MI-DORF to the partially-observed setting, where labels for a small subset of instances are available during training. This scenario has been previously explored using Latent-dynamical models such as Conditional Random Fields [29] and their extensions (HCRF [30]). Although the instance labels are incorporated in these approaches, they can be considered suboptimal for MI-DOR, where sequence weak-labels need to be also taken into account according to the MIL assumptions.

C. Non-Supervised Facial Behavior Analysis

Research on automatic facial behavior analysis has mainly focused on the fully-supervised setting. In the specific problems of Action Unit and Pain Intensity Estimation, recent works have developed models based on HCORF [9], Metric Learning [31], Convolutional Neural Networks [11] or Gaussian Processes [32] among others. However, as discussed in Sec I, supervised models are limited in this context because they involve a laborious data labelling.

In order to reduce the annotation efforts, in this work we address these problems using weakly-supervised learning, which lies on the spectrum in between the unsupervised and fully supervised paradigms. In this context, previous works have explored non-supervised approaches for Facial Behavior Analysis. For AU detection, Zhou *et al.* [33] proposed Aligned Cluster Analysis for the unsupervised segmentation and clustering of facial events in videos. Their experiments showed that the obtained clusters were coherent with AU manual annotations. We find another example in [34], where Multiple Instance Classification was used to find key frames representing Action Unit activations in sequences. Different from these cited approaches which focus on binary detection, we address weakly-supervised Action Unit intensity estimation. To this end, the proposed MI-DORF model is able to learn from segments which are labelled according to the increasing or decreasing evolution of AU intensities (see Sec. I-A). A similar problem has been recently addressed by Zhao *et al.* [35]. Specifically, Ordinal Support Vector Ordinal Regression (OSVR) was used to estimate facial expression intensities using only onset and appex segments during training. However, OSVR presents some limitations in this context. Firstly, it models the instance (frame) labels as continuous variables which is a sub-optimal modelling of ordinal variables. Secondly, OSVR poses MI-DOR as a ranking problem causing the scale of predicted values to not necessarily match with the ground-truth. In contrast, MI-DORF models instance labels as ordinal variables, thus allowing to better estimate labels scale by determining a priory the number of ordinal levels. Finally, OSVR is an static approach and temporal correlations are not modelled as in MI-DORF.

In the context of weakly-supervised Pain Intensity estimation, MIL approaches have been previously applied by considering that a weak-label is provided for a sequence (in terms of the maximum pain intensity felt by the patient).

Then, a video is considered as a bag and image frames as instances. Sikka *et al.* [36] proposed to extract a Bag-of-Words representation from video segments and treat them as bag-instances. Then, MILBoosting [23] was applied to predict sequence-labels under the MIC assumption. Following the bag-based paradigm, [3] developed the Regularized Multi-Concept MIL method capable of discovering different discriminative pain expressions within a sequence. More recently, [14] proposed MI Hidden Markov Models, an adaptation of standard HMM to the MIL problem. The limitation of these approaches is that they focus on the binary detection problem (i.e, pain intensity levels are binarized) and thus, are unable to consider different intensity levels of pain. This is successfully attained by the proposed MI-DORF.

III. Multi-Instance Dynamic Ordinal Regression

In this section, we formalize the MI-DOR problem and its particular instances addressed in this work: Maximum MI-DOR and Relative MI-DOR. In these tasks we are provided with a training set $\mathcal{T} = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \ldots, (\mathbf{X}_N, y_N)\}$ formed by pairs of structured-inputs $X \in \mathcal{X}$ and labels y. Specifically, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ are temporal sequences of T observations $\mathbf{x} \in R^d$ in a d-dimensional space. Given the training-set $\mathcal{T}$, the goal is to learn a model $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{H}$ mapping sequences $\mathbf{X}$ to an structured-output $\mathbf{h} \in \mathcal{H}$. Concretely, $\mathbf{h} = \{h_1, h_2, \ldots, h_T\}$ is a sequence of variables $h_t \in \{0 \prec \ldots \prec l \prec L\}$ assigning one ordinal value for each observation $\mathbf{x_t}$. In order to learn the model $\mathcal{F}$ from $\mathcal{T}$, it is necessary to incorporate prior knowledge defining the Multi-Instance relation between labels y and latent ordinal states $\mathbf{h}$. In Maximum MI-DOR, we assume that bag-labels $y \in \{0 \prec \ldots \prec l \prec L\}$ are also ordinal variables and that the maximum value in $\mathbf{h}_n$ must be equal to y_n:

$$y_n = \max_h(\mathbf{h}_n) \quad \forall\ (\mathbf{X}_n, y_n) \in \mathcal{T} \tag{1}$$

On the other hand, in Relative MI-DOR the sequence label is a categorical variable taking four possible values $y \in \{\uparrow, \downarrow, \emptyset, \updownarrow\}$. Each label indicates the type of evolution within latent labels $\mathbf{h}$. Concretely, in sequences labelled with $y = \uparrow$, there must be an increasing ordinal level transition in, at least, one instant t. Moreover, no decreasing transitions are allowed within the sequence. The opposite occurs in sequences labelled as $y = \downarrow$. In the case of $y = \updownarrow$ the sequence is assumed to contain decreasing and increasing transitions. Finally, when $y = \emptyset$ all the ordinal values in $\mathbf{h}$ should be equal (monotone sequence). Formally, these constraints can be defined as:

$$\forall\ (\mathbf{X}_n, y_n) \begin{cases} y_n = \uparrow & \text{iff} \quad (\exists t\ h_t < h_{t+1}) \wedge (\forall t\ h_t \leq h_{t+1}) \\ y_n = \downarrow & \text{iff} \quad (\exists t\ h_t > h_{t+1}) \wedge (\forall t\ h_t \geq h_{t+1}) \\ y_n = \emptyset & \text{iff} \quad (\forall t\ h_t = h_{t+1}) \\ y_n = \updownarrow & \text{otherwise} \end{cases} \tag{2}$$

Note that the definition of these MI-DOR problems differs from standard supervised sequence classification with latent variables. In that case, the main goal is to learn a model $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ mapping $\mathbf{X}$ to sequence labels y.

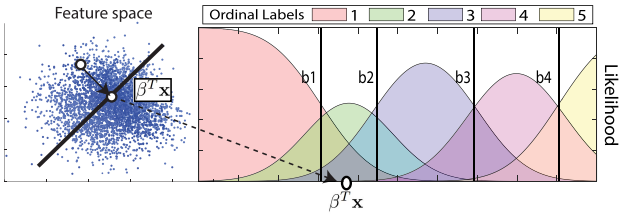

Fig. 2. Illustration of the ordered probit model employed to define the ordinal node potentials in MI-DORF. Vector β is used to project feature vector x onto continuous values. Thresholds β define different bins over the projection determining the likelihood for each ordinal level.

IV. Max-Multi-Instance Dynamic Ordinal Random Fields (MaxMI-DORF)

In this section, we present the proposed Max-Multi-Instance Dynamic Ordinal Random Fields to solve the Maximum MI-DOR problem described in Sec. III.

A. Model Definition

MaxMI-DORF is an Undirected Graphical Model defining the conditional probability of labels y given observations $\mathbf{X}$ with a Gibbs distribution:

$$P(y|\mathbf{X};\theta) = \sum_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{X};\theta) = \frac{\sum_h e^{-\Psi(\mathbf{X},\mathbf{h},y;\theta)}}{\sum_{y'}\sum_h e^{-\Psi(\mathbf{X},\mathbf{h},y';\theta)}}, \tag{3}$$

where θ is the set of the model parameters and the energy function $\Psi(\mathbf{X}, \mathbf{h}, y; \theta)$ is composed of the sum of three different types of potentials (see Fig. 3(a)):

$$\sum_{t=1}^{T} \Psi^N(\mathbf{x}_t, h_t; \theta^N) + \sum_{t=1}^{T-1} \Psi^E(h_t, h_{t+1}; \theta^E) + \Psi^M(\mathbf{h}, y, \theta^M), \tag{4}$$

1) MaxMI-DORF: Ordinal Node Potential: This potential $\Psi^N(\mathbf{x}, h; \theta^N)$ aims to capture the compatibility between a given observation $\mathbf{x}_t$ and the latent ordinal value h_t. Similar to HCORF, it is defined using the ordered probit model [37]:

$$\Psi^N(\mathbf{x}, h = l) = \log\left[\Phi\left(\frac{b_l - \beta^T\mathbf{x}}{\sigma}\right) - \Phi\left(\frac{b_{(l-1)} - \beta^T\mathbf{x}}{\sigma}\right)\right], \tag{5}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function (CDF), and $\theta^N = \{\beta, \mathbf{b}, \sigma\}$ is the set of potential parameters. Specifically, the vector $\beta \in R^d$ projects observations $\mathbf{x}$ onto an ordinal line divided by a set of cut-off points $b_0 = -\infty \leq \cdots \leq b_L = \infty$. Every pair of contiguous cut-off points divide the projection values into different bins corresponding to the different ordinal states $l = 1, \ldots, L$. The difference between the two CDFs provides the probability of the latent state l given the observation $\mathbf{x}$, where σ is the standard deviation of a Gaussian noise contaminating the ideal model (see Fig. 2 and [13] for more details). In our case, we fix $\sigma = 1$, to avoid model over-parametrization. This type of potentials has previously been shown to be effective for Ordinal Regression problems such as AU or Pain Intensity estimation [9], [10].

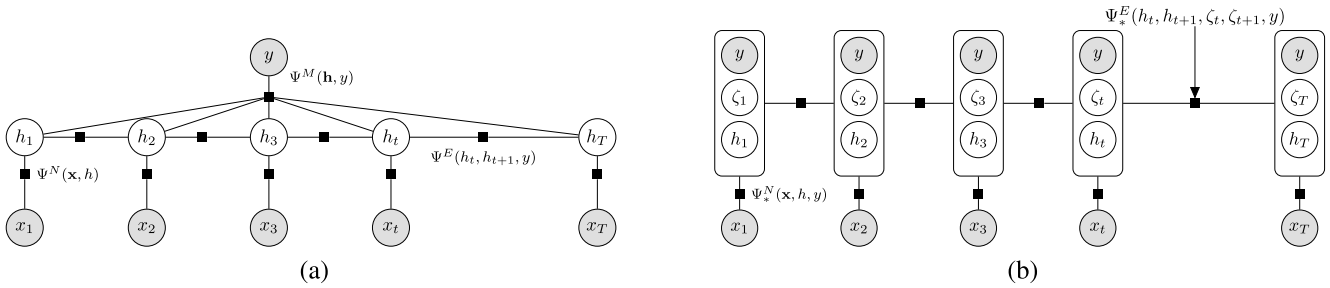

Fig. 3. (a) Factor graph representation of the proposed MI-DORF framework. Node potentials Ψ^N model the compatibility between a given observation $\mathbf{x}_t$ and a latent ordinal value h_t. Edge potentials Ψ^E take into account the transition between consecutive latent ordinal states h_t and h_{t+1}. Finally, the high-order potential Ψ^M models Multi-Instance assumptions relating all the latent ordinal states $\mathbf{h}_t$ with the bag-label y. (b) Equivalent model to MI-DORF defined using the auxiliary variables ζ_t for each latent ordinal state. The use of these auxiliary variables and the redefinition of node and edge potentials allows to perform efficient inference by removing the high-order dependency introduced by the potential Ψ^M (see Sec. IV-C and V-B).

2) MaxMI-DORF: Edge Potentials: The edge potential $\Psi^E(h_t, h_{t+1}; \theta^E)$ models temporal information regarding compatibilities between consecutive latent ordinal states as:

$$\Psi^E(h_t = l, h_{t+1} = l'; \theta^E) = f(\mathbf{W}_{l,l'}), \tag{6}$$

where $\theta^E = \mathbf{W}^{L\times L}$ represents a real-valued transition matrix as in standard HCR. On the other hand, f is a non-linear function defined as $f(s) = -\log(1 + exp(-s))$. The motivation of using f is to maintain the same range between the values of node and edge potentials. Specifically, f bounds the value of Ψ^E between $[0, -\infty]$ as in the case of the node potentials.

3) MaxMI-DORF: Multi-Instance Potential: In order to model the Maximum MI-DOR assumption (see Eq. 1), we define a high-order potential $\Psi^M(\mathbf{h}, y; \theta^M)$ involving label y and all the sequence latent variables $\mathbf{h}$ as:

$$\Psi^M(\mathbf{h}, y; \theta^M) = \begin{cases} w \sum_{t=1}^{T} \mathbf{I}(h_t == y) & \text{iff} \quad \max\ (\mathbf{h}) = y \\ -\infty & \text{otherwise,} \end{cases} \tag{7}$$

where $\mathbf{I}$ is the indicator function, and $\theta^M = w$. Note that when the maximum value within $\mathbf{h}$ is not equal to y, the energy function is equal to $-\infty$ and, thus, the probability $P(y|\mathbf{X}; \theta)$ drops to 0. On the other hand, if the MI assumption is fulfilled, the summation $w \sum_{t=1}^{T} \mathbf{I}(h_t == y)$ increases the energy proportionally to w and the number of latent states $\mathbf{h} \in h_t$ that are equal to y. This is convenient since, in sequences annotated with a particular label, it is more likely to find many latent ordinal states with such ordinal level. Eq. 7 shares some relations with the cardinality potentials [38] also employed in binary Multi-Instance Classification [26].

B. MaxMI-DORF: Learning

Given a training set $\mathcal{T}$, we learn the model parameters θ by minimizing the regularized log-likelihood:

$$\min_{\theta} \quad \sum_{i=1}^{N} \log P(y|\mathbf{X}; \theta) + \mathcal{R}(\theta), \tag{8}$$

where the regularization function $\mathcal{R}(\theta)$ over the model parameters is defined as:

$$\mathcal{R}(\theta) = \alpha\ (||\beta||_2^2 + ||\mathbf{W}||_F^2) \tag{9}$$

and α is set via a validation procedure. We use L2 regularization because, in related Latent Variable models such as HCRF [27] or HCORF [13], it has been shown to provide an effective mechanism to reduce overfitting.

The objective function in Eq.8 is differentiable and standard gradient descent methods can be applied for optimization. To this end, we use the L-BFGS Quasi-Newton method [39]. The gradient evaluation involves marginal probabilities $p(h_t|\mathbf{X})$ and $p(h_t, h_{t+1}|\mathbf{X})$ which can be efficiently computed using the proposed algorithm in Sec. IV-C.

C. MaxMI-DORF: Inference

The evaluation of the conditional probability $P(y|\mathbf{X}; \theta)$ in Eq.3 requires computing $\sum_h e^{-\Psi(\mathbf{X}, \mathbf{h}, y; \theta)}$ for each label y. Given the exponential number of possible latent states $\mathbf{h} \in \mathcal{H}$, efficient inference algorithms need to be used. In the case of Latent-Dynamic Models such as HCRF/HCORF, the forward-backward algorithm [17] can be applied. This is because the pair-wise linear-chain connectivity between latent states $\mathbf{h}$. However, in the case of MaxMI-DORF, the inclusion of the MIL potential $\Psi^M(\mathbf{h}, y; \theta^M)$ introduces a high-order dependence between the label y and all the latent states in $\mathbf{h}$. Inference methods with cardinality potentials have been previously proposed in [38] and [40]. However, these algorithms only consider the case where latent variables are independent and, therefore, they can not be applied in our case. For these reasons, we propose an specific inference method. The idea behind it is to apply the standard forward-backward algorithm by converting the energy function defined in Eq. 4 into an equivalent one preserving the linear-chain connectivity between latent states $\mathbf{h}$.

To this end, we introduce a new set of auxiliary variables $\boldsymbol{\zeta} = \{\zeta_1, \zeta_2, \ldots, \zeta_T\}$, where each $\zeta_t \in \{0, 1\}$ takes a binary value denoting whether the sub-sequence $\mathbf{h}_{1:t}$ contains at least one ordinal state h equal to y. Now we define an alternative MaxMI-DORF energy function Ψ_* as:

$$\Psi_*(\mathbf{X}, \mathbf{h}, \boldsymbol{\zeta}, y; \theta) = \sum_{t=1}^{T} \Psi_*^N(\mathbf{x}_t, h_t, \zeta_t, y; \theta^N) + \sum_{t=1}^{T-1} \Psi_*^E(h_t, h_{t+1}, \zeta_t, \zeta_{t+1}, y; \theta^E), \tag{10}$$

where the new node pontentials Ψ^N_* and edge potentials Ψ^E_* are given by:

$$\Psi^N_* = \begin{cases} \Psi^N(\mathbf{x}_t, h_t; \theta^N) + w\mathbf{I}(h_t = y) & \text{iff } h_t <= y \\ -\infty & \text{otherwise} \end{cases}$$

$$\Psi^E_* = \begin{cases} \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = 0 \wedge \zeta_{t+1} = 0 \wedge h_{t+1} \neq y \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = 0 \wedge \zeta_{t+1} = 1 \wedge h_{t+1} = y \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = 1 \wedge \zeta_{t+1} = 1 \\ -\infty & \text{otherwise} \end{cases} \quad (11)$$

Note that Eq. 10 does not include the potential Ψ^M, thus, the high-order dependence between the label y and latent ordinal-states $\mathbf{h}$ is removed. The graphical representation of MI-DORF with the redefined energy function is illustrated in Fig.3(b). In order to show the equivalence between energies in Eqs. 4 and 10, we explain how the the original Multi-Instance potential Ψ^M is incorporated into the new edge and temporal potentials. Firstly, note that Ψ^N now also takes into account the proportion of ordinal variables h_t that are equal to the sequence label. Moreover, it enforces $\mathbf{h}$ not to contain any h_t greater than y, thus aligning the bag and (max) instance labels. However, the original Multi-Instance potential also constrained $\mathbf{h}$ to contain at least one h_t with the same ordinal value than y. This is achieved by using the set of auxiliary variables ζ_t and the re-defined edge potential Ψ^E. In this case, transitions between latent ordinal states are modelled but also between auxiliary variables ζ_t. Specifically, when the ordinal state in h_{t+1} is equal to y, the sub-sequence $\mathbf{h}_{1:t+1}$ fulfills the Maximum MI-DOR assumption and, thus, ζ_{t+1} is forced to be 1. By defining the special cases at the beginning and the end of the sequence ($t = 1$ and $t = T$):

$$\Psi^N_*(\mathbf{x}_1, h_1, \zeta_1, y) = \begin{cases} \Psi^N_* & \text{iff } \zeta_1 = 0 \wedge l_1 < y \\ \Psi^N_* & \text{iff } \zeta_1 = 1 \wedge l_1 = y \\ -\infty & \text{otherwise,} \end{cases} \quad (12)$$

$$\Psi^N_*(\mathbf{x}_T, h_T, \zeta_T, y) = \begin{cases} \Psi^N_* & \text{iff } \zeta_T = 1 \wedge h_T <= y \\ -\infty & \text{otherwise} \end{cases} \quad (13)$$

we can see that the energy is $-\infty$ when the Maximum MI-DOR assumption is not fulfilled. Otherwise, it has the same value than the one defined in Eq.4 since no additional information is given. The advantage of using this equivalent energy function is that the standard forward-backward algorithm can be applied to efficiently compute the conditional probability:

$$P(y|\mathbf{X}; \theta) = \frac{\sum_{\mathbf{h}} \sum_{\boldsymbol{\zeta}} e^{-\Psi_*(\mathbf{X},\mathbf{h},\boldsymbol{\zeta},y;\theta)}}{\sum_{y'} \sum_{\mathbf{h}} \sum_{\boldsymbol{\zeta}} e^{-\Psi_*(\mathbf{X},\mathbf{h},\boldsymbol{\zeta},y';\theta)}}, \quad (14)$$

The proposed procedure has a computational complexity of $\mathcal{O}(T \cdot (2L)^2)$ compared with $\mathcal{O}(T \cdot L^2)$ using standard forward-backward in traditional linear-chain latent dynamical models. Since typically $L << T$, this can be considered a similar theoretical complexity. The presented algorithm can also be applied to compute the marginal probabilities $p(h_t|\mathbf{X})$ and $p(h_t, h_{t+1}|\mathbf{X})$.

V. Relative-Multi-Instance DORF (RelMI-DORF)

In this section, we present the proposed Relative-Multi-Instance Dynamic Odinal Random Fields to solve the Relative MI-DOR problem described in Sec. III.

A. RelMI-DORF: Model Definition

In RelMI-DORF, ordinal and node potentials are specified as in MaxMi-DORF. However, the Multi-Instance potential $\Psi^M(\mathbf{h}, y)$ it is now defined as shown in Eq. 15. In this case, the potential models the Relaltive MI-DOR assumption, i.e, the weak-relation between the sequence label y and the evolution of latent instance labels $\mathbf{h}$ (see Eq. 2).

$$\Psi^M = \begin{cases} 0 & \text{iff } (\exists\, t h_t < h_{t+1}) \wedge (\forall t\ h_t \leq h_{t+1}) \wedge y = \uparrow \\ 0 & \text{iff } (\exists t\ h_t > h_{t+1}) \wedge (\forall t\ h_t \geq h_{t+1}) \wedge y = \downarrow \\ 0 & \text{iff } (\exists t\ h_t > h_{t+1}) \wedge (\exists t\ h_t < h_{t+1}) \wedge y = \updownarrow \\ 0 & \text{iff } (\forall t\ h_t = h_{t+1}) \wedge y = \emptyset \\ -\infty & \text{otherwise} \end{cases} \quad (15)$$

Learning in RelMI-DORF can be performed following the same procedure described in Sec. IV-B. However, inference requires a special treatment which is described as follows.

B. RelMI-DORF: Inference

Similar to the case of MaxMI-DORF, the high-order potential $\Psi^N(\mathbf{h}, y)$ in RelMI-DORF prevents to perform inference using the standard forward-backward procedure. For this purpose, we follow a similar strategy than the one described in Sec. IV-C. However, in this case, auxiliary variables ζ_t are defined according to the possible sequence labels in Relative MI-DOR. Concretely, $\zeta_t \in \{\uparrow, \downarrow, \emptyset, \updownarrow\}$ indicates the label of the subsequence $\mathbf{h}_{1:t}$ according to the definitions given in Eq. 2. The equivalent energy function incorporating this auxiliary variables ζ can be obtained by redefining the original edge potentials as:

$$\Psi^E_* = \begin{cases} \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \emptyset \wedge \zeta_{t+1} = \emptyset \wedge h_t = h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \emptyset \wedge \zeta_{t+1} = \uparrow \wedge h_t < h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \emptyset \wedge \zeta_{t+1} = \downarrow \wedge h_t > h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \uparrow \wedge \zeta_{t+1} = \uparrow \wedge h_t \leq h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \uparrow \wedge \zeta_{t+1} = \updownarrow \wedge h_t > h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \downarrow \wedge \zeta_{t+1} = \downarrow \wedge h_t \geq h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \downarrow \wedge \zeta_{t+1} = \updownarrow \wedge h_t < h_{t+1} \\ \mathbf{W}_{h_t,h_{(t+1)}} & \text{iff } \zeta_t = \updownarrow \wedge \zeta_{t+1} = \updownarrow \\ -\infty & \text{otherwise} \end{cases} \quad (16)$$

Again, defining the special cases for node potentials at the beginning and ending of the sequence:

$$\Psi^N_*(\mathbf{x}_1, h_1, \zeta_1, y) = \begin{cases} \Psi^N(\mathbf{x}_1, h_1, y) & \text{iff } \zeta_1 = \emptyset \\ -\infty & \text{otherwise,} \end{cases} \quad (17)$$

$$\Psi^N_*(\mathbf{x}_T, h_T, \zeta_T, y) = \begin{cases} \Psi^N(\mathbf{x}_T, h_T, y) & \text{iff } \zeta_T = y \\ -\infty & \text{otherwise,} \end{cases} \quad (18)$$

it can be shown that the energy function becomes $-\infty$ when the sequence level is not coherent with the evolution of latent instance labels $\mathbf{h}$ (according to sequence label y and the Relative MI-DOR assumption). Otherwise, it takes the same value than the energy function defined by the original potentials. In this case, computational complexity is $\mathcal{O}(T \cdot (4L)^2)$, which is still linear in terms of instances T.

VI. Partially-Observed MI-DOR (PoMI-DOR)

Although labels at sequence-level are easier to collect, in some applications is feasible to annotate a small subset of the sequence's instances. In this case, we are interested in learning the model by using weak-labels y but also incorporating the information of these additional annotations. We refer to this problem as Partially-Observed Multi-Instance Dynamic Ordinal Regression (PoMI-DOR). In this case, the training set is formed by triples $\mathcal{T} = \{(\mathbf{X}_1, y_1, \mathbf{h}_1^a), (\mathbf{X}_2, y_2, \mathbf{h}_2^a), \ldots, (\mathbf{X}_N, y_N, \mathbf{h}_N^a)\}$, where $\mathbf{h}_n^a$ contains ground-truth annotations for a subset of sequence instances. Formally, the set $\mathbf{h}_n = \{\mathbf{h}_n^a \cup \mathbf{h}_n^u\}$, where $\mathbf{h}_n^u$ is the subset of ordinal labels corresponding to non annotated instances. Under this setting, we extend MI-DORF to learn a model maximizing the log-likelihood function of the conditional probability:

$$P(y, \mathbf{h}_a | \mathbf{X}; \theta) = \frac{\sum_{\mathbf{h}^u} e^{-\Psi(\mathbf{X}, \mathbf{h}^u, \mathbf{h}^a, y; \theta)}}{\sum_{y'} \sum_{\mathbf{h}^u} \sum_{\mathbf{h}^a} e^{-\Psi(\mathbf{X}, \mathbf{h}^u, \mathbf{h}^a, y'; \theta)}}, \quad (19)$$

for all the sequences in the training set. Note that in this case, the knowledge provided by annotated instances $\mathbf{h}_n^a$ is incorporated into the likelihood function. In order to learn a PoMI-DORF model, the same algorithms presented in Secs. IV and V can be applied. However, during inference we need to take into account annotations $\mathbf{h}_n^a$ for each sequence. This can be easily achieved by redefining the original node potentials in RelMI-DORF and MaxMI-DORF as:

$$\Psi^N(\mathbf{x}, h_t) = \begin{cases} -\infty & \text{iff } (h_t \in \mathbf{h}^a) \wedge (h_t^a \neq h_t) \\ \Psi^N(\mathbf{x}_t, h_t) & \text{otherwise,} \end{cases} \quad (20)$$

Intuitively, observed instance labels $\mathbf{h}^a$ are treated as hard evidences which make the energy function to take a value of $-\infty$ when $\mathbf{h}$ is not consistent with them. This strategy has been previously followed in order to learn Conditional Random Fields [29] under the partially-observed setting.

VII. Experiments

A. Compared Methods

The presented frameworks are designed to address Multi-Instance-Learning problems when bags are structured as temporal sequences of instances with ordinal labels. Given that this has not been attempted before, we evaluate alternative methods that can be also used in these problems but present some limitations: either ignore the MIL assumptions (Single-Instance), do not model dynamic information (Static) or do not take into account the ordinal nature of instance labels.

1) Single-Instance Ordinal Regression (SIL-OR): Maximum MI-DOR can be posed as a supervised learning problem with noisy labels. The main assumption is that the majority of instances will have the same label than their bag. In order to test this assumption, we train standard Ordinal Regression [37] at instance-level by setting all their labels to the same value as their corresponding bag. This baseline can be considered an Static-SIL approach to solve the Maximum MI-DOR problem.

2) Static Multi-Instance Ordinal Regression (MI-OR): Again for Maximum MI-DOR, we have implemented this Static Multi-Instance approach. This method is inspired by MI-SVM [22], where instance labels are considered latent variables and are iteratively optimized during training. To initialize the parameters of the ordinal regressor, we follow the same procedure as described above in SIL-OR. Then, ordinal values for each instance are predicted and modified so that the Maximum MI-DOR assumption is fulfilled for each bag. Ordinal Regression is applied again and this procedure is applied iteratively until convergence.

3) Multi-Instance-Regression (MIR): As discussed in Sec. I, the Maximum MI-DOR problem is closely related with Multiple-Instance-Regression. In order to evaluate the performance of this strategy, we have implemented a similar method as used in [25]. Note that this approach does not model temporal information and treat ordinal labels as continuous variables.

4) MaxMI-DRF: This approach is similar to the proposed MaxMI-DORF. However, MaxMI-DRF ignores the ordinal nature of labels and models them as categorical variables. For this purpose, we replace the MaxMI-DORF node potentials by a multinomial logistic regression model [41].[1] Inference is performed by using the same algorithm described in Sec. IV-C.

5) RelMI-DRF: Similar to MaxMI-DRF, this method is equivalent to RelMI-DORF but modelling instance labels as categorical variables.

6) Latent-Dynamic Models (HCRF/HCORF): In Maximum and Relative MI-DOR a label at sequence-level is provided during training. Therefore, it is possible to apply existing Latent-Dynamic Models such as HCRF [27] or HCORF [13] for both problems. Despite these two methods model dynamics and incorporate the information provided by sequence-labels, they do not take into account the Multi-Instance assumptions.

7) Ordinal Support Vector Regression (OSVR): This method presented in [35] can be applied for Relative MI-DOR. However, it is an Static approach that do not consider dynamic information. Moreover, it models instance labels as continuous variables instead of ordinal.

8) Methods for Partially-Observable MI-DOR: In our experiments, we evaluate Max-MIDORF and Rel-MIDORF when some instance labels are also available during training (see Sec. VI). In order to compare their performance under this setting, we evaluate the partially-observed extensions of CRF [29] and HCRF [30]. Ordinal versions of these two approaches has been also implemented for this work.

[1]The potential with the Multinomial Logistic Regession model is defined as $\log\left(\frac{\exp(\beta_l^T x)}{\sum_{l' \in L} \exp(\beta_{l'}^T x)}\right)$. Where all $\beta_{\mathbf{l}}$ defines a linear projection for each possible ordinal value l [41]

9) Methods for Supervised Dynamic Ordinal Regression: To fully evaluate the performance of methods trained using only weak-labels, we compare the previous described methods with two related fully-supervised models for sequence classification CRF [42] and CORF [43]. These approaches are learned with complete information (i.e, labels for all the instances).

B. Metrics and Evaluation

In order to evaluate the performance of the different methods, we report results in terms of instance-labels predictions. Note that in the MIL literature, results are usually reported at bag-level. However, in MI-DOR problems, the only goal is to predict instance labels (pain or AU intensities) inside the bag (video). Given the ordinal nature of the labels, we use Pearson's Correlation (CORR), Mean-Average-Error (MAE) and Intra-Class-Correlation (ICC) as evaluation metrics. In all our experiments, we used a subset of the training sequences to optimize the different regularization weights (hyper-parameters) in a cross-validation procedure. To this end, we used standard grid-search where regularization parameters has been chosen between different values in the range $[10^{-4}, 10^{-1}]$.

C. Maximum MI-DOR and Relative MI-DOR: Synthetic Data

1) Synthetic Data Generation: Given that no standard benchmarks are available for MI-DOR problems, we have generated synthetic data. In order to create sequences for Maximum MI-DOR, we firstly sample a sequence of ordinal values using a random transition matrix representing transition probabilities between temporally-consecutive ordinal levels. Secondly, we generate random parameters of an Ordinal Regressor as defined in Eq. 5. This regressor is used to compute the probabilities for each ordinal level in a set of feature-vectors randomly sampled from a Gaussian distribution. Thirdly, the corresponding sequence observation for each latent state in the sequence is randomly chosen between the sampled feature vectors according to the obtained probability for each ordinal value. Finally, the sequence-label is set to the maximum ordinal state within the sequence following the Maximum MI-DOR assumption and Gaussian noise ($\sigma = 0.25$) is added to the feature vectors. Fig. 4(a-c) illustrates this procedure.

For Relative MI-DOR, we follow a similar strategy to generate the synthetic sequences. However, the transition matrix is forced to contain a probability of 0 for decreasing transitions in case the sequence label is $y = \uparrow$ and for increasing transitions if $y = \downarrow$. For testing, we create unsegmented sequences (with increasing and decreasing transitions) by concatenating two segments generated following the previous procedure.

2) Experimental Setup and Results: Following the strategy described above, we have generated ten different data sets for Relative and Maximum MI-DOR by varying the ordinal regressor parameters and transition matrix. Specifically, each dataset is composed of 100 sequences for training, 150 for testing and 50 for validation. The sequences have a variable length between 50 and 75 instances in Maximum MI-DOR and between 15 and 25 in Relative MI-DOR. The dimensionality of the feature vectors was set to 10 and the number of ordinal values to 6. For partially-observed MI-DOR, we have randomly choose one instance per sequence of which its label is also used during training. Table I and II shows the results computed as the average performance over the ten datasets for Maximum and Relative MI-DOR respectively. We also

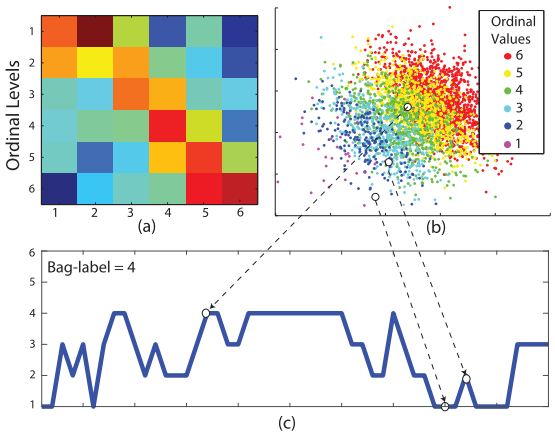

Fig. 4. Description of the procedure used to generate synthetic sequences. (a) A random matrix modelling transition probabilities between consecutive latent ordinal values. (b) Ordinal levels assigned to the random feature vectors according to the ordinal regressor. (c) Example of a sequence of ordinal values obtained using the generated transition matrix. The feature vector representing each observation is randomly chosen between the samples in (b) according to the probability for each ordinal level.

TABLE I
RESULTS ON SYNTHTIC DATA (MAXMI-DOR)

Setting	Method	CORR ↑	MAE ↓	ICC ↑
MaxMI-DOR	SI-OR	0.79	1.31	0.46
	MI-OR	0.82	0.62	0.70
	HCRF [27]	0.05	1.99	0.05
	HCORF [13]	0.73	0.74	0.65
	MIR [25]	0.79	0.65	0.69
	MaxMI-DRF	0.77	0.77	0.71
	MaxMI-DORF	**0.86**	**0.41**	**0.85**
PoMaxMI-DOR (1 sample/seq.)	PoCRF [29]	0.74	0.63	0.74
	PoCORF [29]*	0.84	0.46	0.83
	PoHCRF [30]	0.79	0.57	0.78
	PoHCORF [30]*	0.86	0.42	0.85
	MaxMI-DRF	0.82	0.52	0.81
	MaxMI-DORF	**0.87**	**0.38**	**0.87**
Supervised DOR	CRF [43]	0.88	0.35	0.88
	CORF [44]	**0.89**	**0.35**	**0.88**

(*)Indicates a nominal method that we have extended to the ordinal case.

TABLE II
RESULTS ON SYNTHTIC DATA (RELMI-DOR)

Setting	Method	CORR ↑	MAE ↓	ICC ↑
RelMI-DOR	HCRF [27]	0.36	1.82	0.32
	HCORF [13]	0.85	1.32	0.80
	OSVR [42]	0.87	3.51	0.10
	RelMI-DRF	0.77	1.36	0.49
	RelMI-DORF	**0.89**	**0.74**	**0.84**
PoRelMI-DOR (1 sample/seq.)	PoCRF [29]	0.82	0.64	0.81
	PoCORF [29]*	0.89	0.43	0.89
	PoHCRF [30]	0.83	0.60	0.83
	PoHCORF [30]*	0.89	0.44	0.88
	OSVR [42]	0.87	0.61	0.85
	RelMI-DRF	0.88	0.49	0.87
	RelMI-DORF	**0.92**	**0.36**	**0.91**
Supervised DOR	CRF [43]	0.93	0.31	0.93
	CORF [44]	**0.93**	**0.29**	**0.93**

(*)Indicates a nominal method that we have extended to the ordinal case.

report results for fully-supervised CRF and CORF trained considering all the instance labels.

3) Maximum MI-DOR Discussion: In the Maximum MI-DOR problem, SIL methods (SIL-OR, HCRF and HCORF) obtain lower performance than their corresponding MIL versions (MI-OR, MaxMI-DRF and MaxMI-DORF) in all the evaluated metrics. This is expected since SIL approaches ignore the Multi-Instance assumption. Moreover, HCORF and MaxMI-DORF obtain better performance compared to HCRF and MaxMI-DRF. This is because the former model instance labels as nominal variables, thus, ignoring their ordinal nature. Finally, note that MaxMI-DORF outperforms the static methods MI-OR and MIR. Although these approaches use the Multi-Instance assumption and incorporate the labels ordering, they do not take into account temporal information. In contrast, MaxMI-DORF is able to model the dynamics of latent ordinal states and use this information to make better predictions when sequence observations are noisy.

Looking into the results achieved by the different methods in the PoMI-DOR setting, we can derive the following conclusions. Firstly, HCORF and HCRF improve their performance by taking into account the additional information provided by instance labels. However, we can observe that, under this setting, CRF and CORF obtain lower results than HCORF and HCRF. This is because the later are able to use the sequence-label information together with the provided by labelled instances. Secondly, observe that MaxMI-DRF and MaxMI-DORF still achieves better performance than methods that do not consider the MIL assumption (CORF, CRF, HCRF and HCORF). This shows the importance of explicitly incorporate the Maximum MI-DOR assumption in the model even though instance labels can be available during training. Finally, note that MaxMI-DORF obtain again the best performance, even close to fully-supervised CRF and CORF. This suggest that the need of annotated instances is highly-reduced if the sequence weak-labels are used during learning.

4) Relative MI-DOR Discussion: In the Relative MI-DOR problem, we observe similar results as in Maximum MI-DOR. Firstly, note that non-ordinal approaches (HCRF and RelMI-DRF) obtain the worst performance in most cases. Secondly, RelMI-DORF obtain better performance than HCORF by explicitly modelling the Multi-Instance-Assumption. Finally, OSVR achieves a competitive performance in terms of correlation compared with RelMI-DORF. However, it obtains poor results in terms of MAE and ICC. As discussed in Sec. II, OSVR considers labels as continuous variables and do not explicitly model the Relative MI-DOR assumption. Instead, it only ranks the instance labels within the sequence. Therefore, it fails to estimate the actual scale of the predicted values.

When some instance labels are provided (PoRel-MIDOR), all the methods improve their performance by exploiting this additional information. However, the improvement in terms of MAE and ICC is much higher than for correlation. This is because in Relative MI-DOR, sequence labels only provide information about the evolution of instance labels within the sequence. Therefore, models can achieve a good performance predicting sequence-labels even though the ordinal levels are not accurate. In contrast, when some instance labels are incorporated during training, a better estimation of the ordinal levels can be achieved. Finally, note that RelMI-DORF under the PoRelMI-DOR setting achieves again competitive performance compared to fully-supervised CRF and CORF.

5) Computational Cost: In order to show the efficiency of the proposed inference algorithms for MaxMI-DORF (Sec. IV-C) and RelMI-DORF (Sec. V-B), we have computed the average time required to process the testing sequences in each of the 10 synthetic datasets used in our experiments.[2] Comparing it with the time required by the forward-backward procedure employed in HCRF and HCORF, MaxMI-DORF is only 1.6 times slower (0.12s vs. 0.08s). Similarly, the forward-backward algorithm is only 1.5 times faster than RelMI-DORF (0.10s vs. 0.07s). Note that the efficiency of the proposed algorithms is better than expected according to our theoretical analysis. This is because our implementation has been optimized by exploiting the inherent sparsity of auxiliary node and edge potentials ($-\infty$ cases in Eq. 11 and 16).

D. Weakly-Supervised Pain Intensity Estimation

In this experiment, we test the performance of MaxMI-DORF for weakly-supervised pain intensity estimation. As detailed in Sec. I-A, our main motivation is that pain intensity labelling is very time consuming. However, the maximum pain felt during a sequence is much easier to annotate.

1) UNBC Dataset: We use the UNBC Shoulder-Pain Database [8] which contains recordings of different subjects performing active and passive arm movements during rehabilitation sessions. In this dataset, pain intensities at each frame are given in terms of the PSPI scale [44]. This ordinal scale ranges from 0 to 15. Given the imbalance between low and high pain intensity levels, we follow the same strategy than [9]. Specifically, pain labels are grouped into 5 ordinal levels as: 0(0),1(1),2(2),3(3),4-5(4),6-15(5). These frame-by-frame pain annotations are considered the instance labels in Maximum MI-DOR. On the other hand, bag (video) labels are extracted as the maximum pain level within each sequence.

In order to extract facial-descriptors at each video frame representing the bag instances, we compute a geometry-based facial-descriptor as follows. Firstly, we obtain a set of 49 landmark facial-points with the method described in [45]. Then, the obtained points are aligned with a mean-shape using Procrustes Analysis. Finally, the facial descriptor are obtained by concatenating the x and y coordinates of the aligned points.

2) Experimental Setup and Results: Similar to the experiment with synthetic data (Sec. VII-C.4), we consider two scenarios for weakly-supervised pain intensity estimation. The first one is the Maximum MI-DOR setting, where only bag labels are used. Apart from the baselines described in Sec. VII-A, in this scenario we also evaluate the performance of the approach presented in [36] which considers pain levels as binary variables. For this purpose, we use the

[2]Average computed over 50 different runs for each dataset. Experiment performed using a MATLAB implementation over a Desktop PC (Intel Core i7-4790K@4.00Ghz processor).

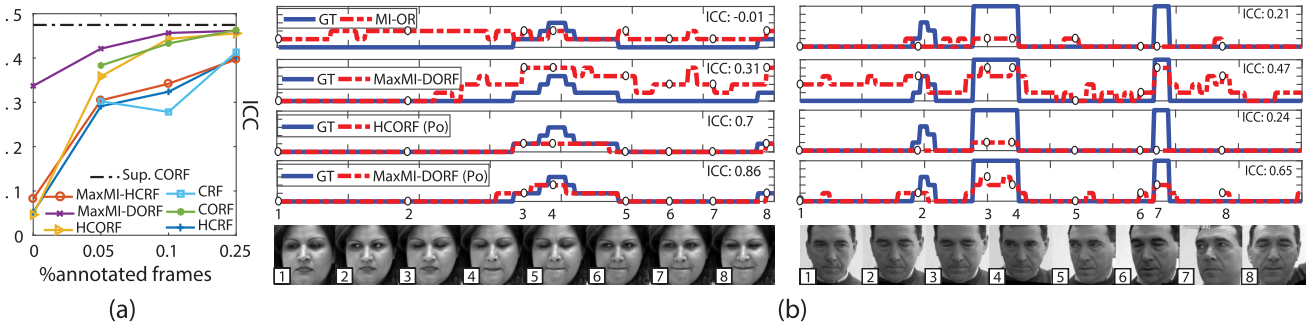

Fig. 5. (a) ICC obtained on the UNBC data when using different percentages of labelled instances from the training set. Black line shows the performance of a fully-supervised CORF trained with all the instance labels. (b) Visualization of the pain intensity predictions in different sequences of the UNBC dataset. From top to bottom: MI-OR and MaxMI-DORF without using instance labels. Partially-observed HCORF and MaxMI-DORF using 10% of annotated frames.

TABLE III
RESULTS ON THE UNBC DATABASE

Setting	Method	CORR ↑	MAE ↓	ICC ↑
MaxMI-DOR	**SI-OR**	0.22	2.20	0.08
	MI-OR	0.29	0.84	0.27
	MILBoost [23]	0.23	2.38	0.09
	HCRF [27]	0.09	1.73	0.05
	HCORF[13]	0.06	1.23	0.05
	MIR [25]	0.32	1.03	0.25
	MaxMI-DRF	0.16	1.96	0.08
	MaxMI-DORF	**0.36**	**0.71**	**0.34**
PoMaxMI-DOR (5% of data)	**PoCRF [29]**	0.31	0.66	0.30
	PoCORF [29]*	0.39	0.58	0.38
	PoHCRF [30]	0.32	0.76	0.29
	PoHCORF [30]*	0.38	0.68	0.36
	MaxMI-DRF	0.32	0.72	0.30
	MaxMI-DORF	**0.43**	**0.52**	**0.42**
PoMaxMI-DOR (10% of data)	**PoCRF [29]**	0.29	0.65	0.28
	PoCORF [29]*	0.44	0.55	0.43
	PoHCRF [30]	0.34	0.63	0.32
	PoHCORF [30]*	0.45	0.58	0.44
	MaxMI-DRF	0.34	0.55	0.34
	MaxMI-DORF	**0.46**	**0.51**	**0.46**
Supervised DOR	**CRF [43]**	0.45	**0.50**	0.44
	CORF [44]	**0.48**	0.56	**0.48**

(*)Indicates a nominal method that we have extended to the ordinal case.

MILBoosting [23] method employed in the cited work and considered videos with a pain label greater than 0 as positive. Given that MI-Classification methods are only able to make binary predictions, we use the output probability as indicator of intensity levels, i.e., the output probability is normalized between 0 and 5.

We also consider the Partially-Observed setting, where different percentages of annotated frames inside each sequence are also available during training. This simulates that the time required to annotate the dataset has been significantly reduced by only labelling a small subset of the frames. Concretely, we consider the 5% and 10% of annotated frames in each sequence. Under these different experimental setups, we perform Leave-One-Subject-Out Cross Validation where, in each cycle, we use 15 subjects for training, 1 for testing and 9 for validation. In order to reduce computational complexity and redundant information between temporal consecutive frames, we have down-sampled the sequences using a time-step of 0.25 seconds. Table III shows the results obtained by the evaluated methods following the described procedure. Results for fully-supervised CRF and CORF are also reported.

3) Discussion: By looking into the results in the Maximum MI-DOR setting, we can derive the following conclusions. Firstly, SI approaches (SI-OR, HCORF and HCRF) obtain worse performance than MI-OR and MIR. Specially, HCORF and HCRF obtain poor results. This is because pain events are typically very sparse in these sequences and most frames have intensity level 0 (neutral). Therefore, the use of the MIL assumption has a critical importance in this problem in order to correctly locate pain frames. Secondly, MIR and MI-OR obtain better results than MaxMI-DRF. This can be explained because the latter consider pain levels as nominal variables and is ignorant of the ordering information of the different pain intensities. Finally, MILBoost trained with binary labels also obtains low performance compared to the MI-OR and MIR. This suggest that current approaches posing weakly-supervised pain detection as a MI-Classification problem are unable to predict accurately the target pain intensities. By contrast, MaxMI-DORF obtains the best performance across all the evaluated metrics. We attribute this to the fact it models the MIL assumption with ordinal variables. Moreover, the improvement of MaxMI-DORF compared to static approaches, such as MI-OR and MIR, suggests that modelling dynamic information is beneficial in this task.

In the Partially-observed setting, all the methods improve their performance by considering the additional information provided by labelled instances. However, note that approaches modelling the ordinal structure of labels (CORF, HCORF and MaxMI-DORF) still outperforms nominal methods (CRF, HCRF and MaxMI-DRF) under this setting. Moreover, MaxMI-DORF also achieves the best performance with 5% and 10% of labeled frames. Despite the other approaches also consider instance labels, MaxMI-DORF better exploits sequence labels information by explicitly modelling the MIL assumption. It is worth mentioning that considering only 10% of annotated frames, MaxMI-DORF obtain competitive performance against fully-supervised approaches. Concretely, it outperforms CRF in terms of ICC/CORR and CORF in terms of MAE. This suggest that the effort needed to annotate pain intensity databases, could be highly-reduced using the proposed weakly-supervised framework. In order to give more insights about this issue, Fig. 5(b) shows the performance in terms of ICC as the percentage of annotated frames increases. As we can observe, MaxMI-DORF outperforms other methods with 0%, 5% and 10% of annotated frames. When this percentage increases to 25%, the performance

of partially-observed CORF, HCORF and MaxMI-DORF is comparable to the achieved by fully-supervised CORF. However, note that labelling 25% of samples does not suppose a significant reduction of the annotation time in a real scenario.

Finally, in Fig. 5(b) we show qualitative examples comparing predictions of the best evaluated methods under the different settings. When only bag-labels are used for training, MI-OR predictions are less accurate than the obtained by MaxMI-DORF. Moreover, MaxMI-DORF estimates better the actual pain levels in the partially-observed setting, where a small subset of instance labels are used. These predictions are more accurate than the obtained with partially-observed HCORF which does not take into account the MIL assumption. This is reflected by the ICC depicted in the sequences, showing that the proposed MaxMI-DORF method outperforms the competing approaches on target data.

E. Weakly-Supervised AU Intensity Estimation

In this section, we test the performance of RelMI-DORF for weakly-supervised Action Unit intensity estimation. Similarly to pain intensity, AU labelling requires a huge effort for expert coders. However, segmenting videos according to the increasing or decreasing evolution of AU intensities (i.e. onset and offset sequences) is less time-consuming.

1) DISFA Dataset: We employ the DISFA Database [7], which is a popular benchmark for AU intensity estimation. It contains naturalistic data consisting on 27 annotated sequences of different subjects watching videos eliciting different types of emotions. Specifically frame-by-frame AU intensities are provided for 12 AUs (1,2,4,5,6,9,12,15,17,20,25,26) in a six-point ordinal scale (*neutral*<A<B<C<D<E). As far as we know, this is the largest available dataset in terms of the number of Action Units annotated. Although the UNBC dataset also provides AU intensity annotations for 11 AUs, we found that the number of onset and appex events for each of them is very limited. Therefore, we discard it for this experiments. To the best of our knowledge, no previous works have evaluated DISFA under the weakly-supervised setting.

The described AU intensities represent the instance labels in our Relative MI-DOR problem. As previously discussed, bags are considered onset and apex sequences where the intensity of a given AU is monotone increasing ($y = \uparrow$) or decreasing ($y = \downarrow$). These segments has been automatically extracted with an exhaustive search over the whole video using the ground-truth intensity labels at frame-level. This procedure simulates that a given annotator has only labelled onset and offset segments instead of specific AU intensities for all the frames. The number of extracted segments for each AU is indicated in Table V. To compute the facial descriptors at each frame, we use the same procedure described in Sec. VII-D.1.

2) Experimental Setup and Results: Using the segments for each AU, we evaluate the different methods using a subject-independent 5-fold cross validation. Specifically, 3 folds are used for training and 1 for testing and validation purposes. During testing, the trained models are evaluated on the original non-segmented videos. The motivation is that, in a real scenario, onset and apex segmentation is not known for testing sequences. We also consider the partially-observed setting, where labels for 5% and 10% of frames are available during training (PoRelMI-DOR). Table IV shows the performance obtained by the evaluated methods computed as the average for all the considered AUs. Specific results in terms of ICC for independent AUs are shown in Table V.

TABLE IV
AVERAGE PERFORMANCE ACROSS AUS ON THE DISFA DATASET

Setting	Method	CORR ↑	MAE ↓	ICC ↑
RelMI-DOR	HCRF [27]	0.21	2.04	0.10
	HCORF [13]	0.26	3.49	0.03
	OSVR [42]	0.35	1.38	0.15
	RelMI-DRF	0.19	1.70	0.11
	RelMI-DORF	**0.40**	**1.13**	**0.26**
PoRelMI-DOR (5% frames)	PoCRF [29]	0.33	0.55	0.29
	PoCORF [29]*	0.37	0.57	0.32
	PoHCRF [30]	0.34	0.59	0.30
	PoHCORF [30]*	0.38	0.62	0.33
	OSVR [42]	0.36	0.81	0.29
	RelMI-DRF	0.23	0.64	0.19
	RelMI-DORF	**0.40**	**0.51**	**0.36**
PoRelMI-DOR (10% frames)	PoCRF [29]	0.36	0.50	0.32
	PoCORF [29]*	0.39	0.56	0.33
	PoHCRF [30]	0.38	0.57	0.34
	PoHCORF [30]*	0.40	0.59	0.35
	OSVR [42]	0.37	0.80	0.29
	RelMI-DRF	0.36	0.50	0.32
	RelMI-DORF	**0.42**	**0.48**	**0.38**
Supervised DOR	CRF [43]	0.39	**0.44**	0.35
	CORF [44]	**0.41**	0.50	**0.37**

(*)Indicates a nominal method that we have extended to the ordinal case.

3) Discussion: When instance labels are not used during training (Relative MI-DOR setting), we can observe that HCRF and HCORF obtain poor results compared to OSVR and RelMI-DORF. This can be explained because the former methods explicitly model the increasing/decreasing intensity constraints provided by sequence weak-labels. Moreover, the low results obtained by RelMI-DRF compared to RelMI-DORF suggest that modelling intensities as nominal variables is suboptimal in this scenario. Also note that OSVR obtains worse results in terms of ICC and MAE compared to RelMI-DORF. Given that performances in terms of CORR are more similar, it shows the limitation of OSVR to predict the actual scale of instance ordinal labels. Considering the results for independent AUs, we observe that RelMI-DORF achieves the best performance for most cases. Note however, that results for some particular AUs (9,15,17, 20) is low for all the methods. We attribute this to the fact that, the activation of these AUs is typically more subtle and high-intensity levels are scarce.

By looking into the results in the partially-observed setting, we can derive the following conclusions. Firstly, all the methods improve their average performance as the percentage of instance labels increases. However, this improvement is more significant for ICC and MAE. This shows that, when instance labels are not available during training, the tendency of intensity levels can be captured. However, accurate predictions of particular ordinal labels requires the additional information provided by frame-by-frame annotations. To illustrate this, in Fig. 6 we show AU12 predictions attained by RelMI-DORF using different percentages of annotated frames. Secondly, note that approaches modelling the ordinal structure of labels usually achieves better performance than nominal methods in terms of ICC and CORR. In contrast, CRF and HCRF

TABLE V

RESULTS (ICC) FOR INDEPENDENT AUS IN THE DISFA DATABASE. IN PARENTHESES, NUMBER OF ONSET AND APEX SEGMENTS EXTRACTED

Setting	Method	AU1 (342)	AU2 (230)	AU4 (572)	AU5 (216)	AU6 (364)	AU9 (159)	AU12 (642)	AU15 (210)	AU17 (575)	AU20 (199)	AU25 (800)	AU26 (723)	AVG
RelMI-DOR	HCRF [27]	0.06	0.03	0.11	0.01	0.03	0.02	0.14	0.01	0.06	0.01	0.45	**0.24**	0.10
	HCORF [13]	0.02	0.01	0.05	0.08	0.02	0.01	0.04	0.01	0.01	0.00	0.06	0.01	0.03
	OSVR [42]	0.10	0.13	0.21	0.04	0.16	0.09	0.40	**0.09**	0.04	0.04	0.37	0.17	0.15
	RMI-HCRF	0.02	0.04	0.10	0.03	0.12	0.01	0.30	0.04	-0.02	0.02	0.40	0.22	0.11
	RMI-DORF	**0.34**	**0.30**	**0.27**	**0.17**	**0.30**	**0.10**	**0.60**	0.07	**0.08**	**0.04**	**0.70**	0.21	**0.26**
PoRelMI-DOR (5% of frames)	PoCRF [29]	0.24	0.33	0.18	0.17	0.40	0.07	0.71	0.14	0.13	0.08	0.85	0.21	0.29
	PoCORF [29]*	0.20	0.39	0.21	0.26	0.41	0.10	0.77	0.14	**0.15**	**0.11**	0.80	0.32	0.32
	PoHCRF [30]	0.26	0.35	0.18	0.17	0.42	0.08	0.72	0.10	0.13	0.08	**0.86**	0.23	0.30
	PoHCORF [30]*	0.24	0.34	0.25	**0.30**	0.40	0.10	**0.78**	0.15	**0.15**	**0.11**	0.81	0.35	0.33
	OSVR [42]	0.15	0.20	**0.30**	0.16	0.34	0.11	0.73	0.16	0.09	**0.09**	0.78	**0.37**	0.29
	RMI-HCRF	0.12	0.39	0.04	0.18	**0.51**	0.10	0.24	0.17	0.06	0.09	0.25	0.14	0.19
	RMI-DORF	**0.38**	**0.47**	0.28	0.29	0.44	**0.11**	**0.78**	**0.18**	**0.15**	**0.11**	0.78	0.35	**0.36**
PoRelMI-DOR (10% of frames)	PoCRF [29]	0.27	0.44	0.21	0.19	0.46	0.06	0.72	**0.22**	0.16	0.07	**0.84**	0.23	0.32
	PoCORF [29]*	0.26	0.45	0.28	0.32	0.39	0.11	0.76	0.17	0.09	0.09	0.78	0.31	0.33
	PoHCRF [30]	0.36	0.46	0.20	0.24	0.40	0.08	0.73	0.26	0.12	0.08	0.84	0.29	0.34
	PoHCORF [30]*	0.25	0.44	0.26	0.35	0.42	0.11	0.77	0.20	0.16	0.09	0.78	0.32	0.35
	OSVR [42]	0.15	0.22	0.29	0.17	0.34	**0.13**	0.74	0.17	0.10	0.09	0.77	**0.37**	0.29
	RMI-HCRF	0.28	0.44	0.24	0.21	**0.49**	0.08	0.71	0.20	0.14	0.12	0.72	0.22	0.32
	RMI-DORF	**0.39**	**0.50**	**0.29**	**0.39**	0.44	0.12	**0.78**	0.21	**0.17**	**0.11**	0.81	0.32	**0.38**
Supervised DOR	CRF [43]	0.33	0.44	0.26	0.33	**0.51**	0.08	0.74	**0.24**	**0.14**	**0.11**	**0.84**	0.24	0.35
	CORF [44]	**0.40**	**0.47**	**0.28**	**0.35**	0.45	**0.11**	**0.78**	0.20	**0.14**	0.09	0.81	**0.32**	**0.37**

(*)Indicates a nominal method that we have extended to the ordinal case.

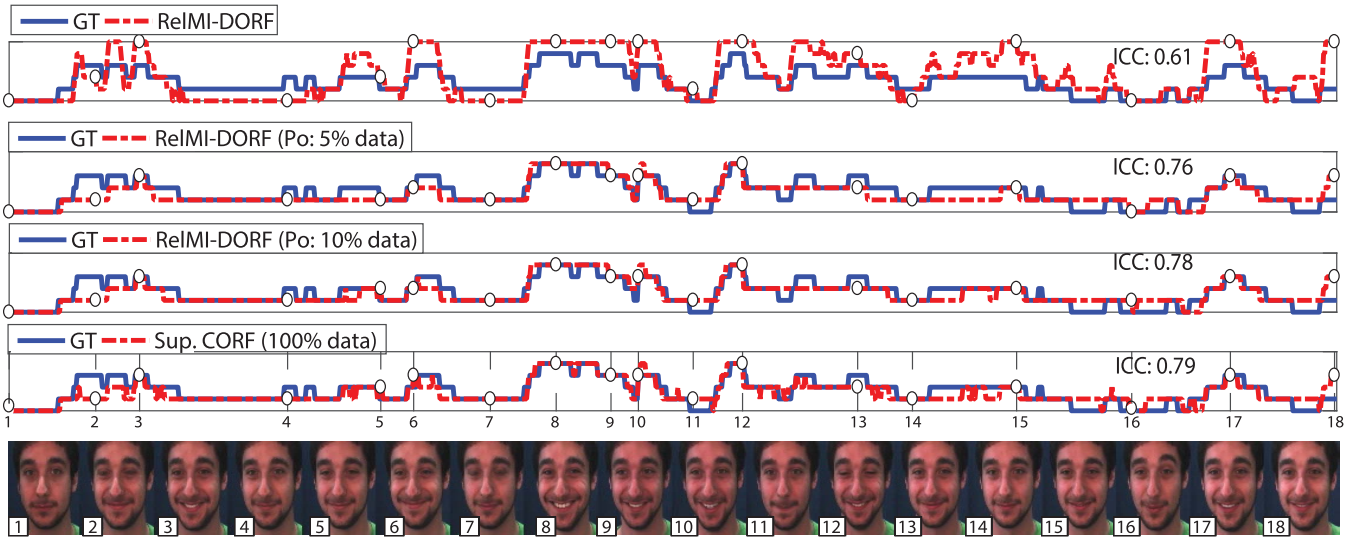

Fig. 6. Visualization of AU12 (Lip-Corner puller) intensity predictions in a subsequence of the DISFA dataset. From top to bottom: RelMI-DORF without using instance labels and with 5% and 10% of annotated frames. Supervised CORF using all the frame labels during training. Intensity estimation for RelMI-DORF tends to be more accurate as more instance labels are considered during training. Using only a 10% of annotated frames, RelMI-DORF achieves similar accuracy than a fully-supervised CORF.

obtain lower MAE than CORF and HCORF. This can be explained because the majority of sequence frames has AU intensity level of 0 (neutral). As a consequence, CRF and HCRF tends to assign most of the frames to this level, thus minimizing the absolute error. In contrast, ordinal methods are more robust to imbalanced intensity levels and capture better changes in AU intensities. Finally, note that the proposed RelMI-DORF method obtain the best average performance considering 5% and 10% of annotated frames. Regarding specific AUs, RelMI-DORF obtain better results for most cases and competitive performance against the best method otherwise. Finally, note that RelMI-DORF performance with 10% of annotated frames is comparable to the achieved by the fully-supervised approaches CRF and CORF. Specifically, only supervised CRF outperforms RelMI-DORF in terms of average MAE. The slightly worse results of supervised CORF compared with RelMI-DORF suggest that considering intensity annotations for all the frames may cause overfitting and decrease performance on unseen test sequences. This can be seen more clearly by looking at the results of independent AUs, where RelMI-DORF obtain slightly better performance than fully-supervised CORF in some cases. In conclusion, the presented results support our hypothesis that it is possible to use the proposed RelMI-DORF model in order to reduce the annotation effort required for AU intensity estimation.

VIII. CONCLUSIONS AND DISCUSSION

In this work, we have presented MI-DORF for the novel task of Multi-Instance Dynamic-Ordinal Regression. To the best of our knowledge, this is the first MIL approach that

imposes an ordinal structure on instance labels, and also attains dynamic modeling within bag instances. By considering different weak-relations between instance and bag labels, we have developed two variants of this framework: RelMI-DORF and MaxMI-DORF. Moreover, we have extended the proposed framework for Partially-Observed MI-DOR problems, where a subset of instance labels are also available during training. Although the presented MI-DORF framework has many potential applications in multiple domains, our results in the context of weakly-supervised facial behavior analysis are relevant in several aspects. In the MI-DOR setting, where no instance-level annotations are available during training, we showed that the proposed method can learn underlying variables that are significantly correlated with the ground-truth instance labels. Even though our results in this setting are lower than fully-supervised approaches, our method provides a good trade-off between the annotation effort and the accuracy of intensity predictions. While we do not claim to replace the AU/Pain annotation process using only weak-labels at sequence-level, this setting may be preferable in some applications. For example, when the focus is on capturing the variation in target facial behaviour rather than obtaining highly accurate frame labels (e.g., for monitoring changes in patient's pain intensity levels), our approach has clear advantages over the fully supervised methods which require a time-consuming annotation process. On the other hand, the competitive results of Partially-Observed MI-DORF compared to the evaluated fully-supervised approaches, indicate that annotation effort can be highly-reduced when combined with weak-information.

It is also worth mentioning recent works on Deep Learning for Action Unit detection [46] and Intensity Estimation [47], [11]. Although these models have a high modelling power, the reported results have not shown significant improvements compared to traditional shallow methods using hand-crafted features. For example, the recently proposed Copula Convolutional Neural Network (CNN) [11] for AU Intensity Estimation is highly-related to our approach, because it combines a CNN with a probabilistic graphical model similar to the one employed in MI-DORF. Even though the Copula CNN requires intensity labels for all the frames during training, the reported results on the DISFA dataset are comparable to those achieved by our method. Specifically, MI-DORF trained with only a 10% of annotated frames obtains better average performance in terms of Mean Average Error (0.48 vs. 0.61) whereas it is outperformed in terms of ICC (0.45 vs. 0.38) (Table IV). Although these results are not directly comparable because of different experimental settings, they indicate that our method trained with labels at sequence-level and a small portion of labelled frames can still show competitive performance. It is known that Supervised Deep Learning models require a large number of samples to be effectively trained [48]. Thus, this still limits their application to Facial Behavior Analysis, where the annotation process is laborious and labelled data is scarce. Posing the facial expression intensity estimation as a weakly-supervised learning problem would provide an opportunity to replace the limited-size datasets currently used in the field, by large-scale not-fully labelled databases. Therefore, coupling Deep models with the proposed framework is a natural step forward and will be the focus of our future research. This would provide a principled way to train these powerful models by taking advantage of data-driven MIL assumptions and a vast amount of weakly-annotated data.

REFERENCES

[1] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[2] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu, "Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 256–263.

[3] A. Ruiz, J. Van de Weijer, and X. Binefa, "Regularized multi-concept MIL for weakly-supervised facial behavior categorization," in *Proc. Brit. Mach. Vis. Conf.*, 2014.

[4] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 1998, pp. 570–576.

[5] S. Ray and D. Page, "Multiple instance regression," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 1–8.

[6] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. London, U.K.: Oxford Univ. Press, 1997.

[7] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 151–160, Apr./Jun. 2013.

[8] 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 *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2011, pp. 57–64.

[9] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 944–958, May 2015.

[10] O. Rudovic and V. Pavlovic, "Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields," in *Proc. Int. Symp. Vis. Comput.*, 2013, pp. 234–243.

[11] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic, "Deep structured learning for facial action unit intensity estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 5709–5718.

[12] A. Ruiz, J. Van de Weijer, and X. Binefa, "From emotions to action units with hidden and semi-hidden-task learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3703–3711.

[13] M. Kim and V. Pavlovic, "Hidden conditional ordinal random fields for sequence classification," *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2010.

[14] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden Markov model for facial expression recognition," in *Proc. IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.*, May 2015, pp. 1–6.

[15] J. Liu, C. Chen, Y. Zhu, W. Liu, and D. N. Metaxas, "Video classification via weakly supervised sequence modeling," *Comput. Vis. Image Understand.*, vol. 152, pp. 79–87, Nov. 2016.

[16] A. Ruiz, O. Rudovic, X. Binefa, and M. Pantic, "Multi-instance dynamic ordinal random fields for weakly-supervised pain intensity estimation," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 171–186.

[17] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[18] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[19] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *Proc. Int. Conf. Mach. Learn.*, 2002, pp. 179–186.

[20] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[21] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-IID samples," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Jun. 2009, pp. 1249–1256.

[22] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. 15th Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 577–584.

[23] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Proc. Adv. Neural Inf. Syst.*, 2005, pp. 1–8.

[24] M. Kim and F. De La Torre, "Gaussian processes multiple-instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 1–8.
[25] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Augmented multiple instance regression for inferring object contours in bounding boxes," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1722–1736, Apr. 2014.
[26] H. Hajimirsadeghi, J. Li, G. Mori, M. Zaki, and T. Sayed, "Multiple instance learning by discriminative training of Markov networks," in *Proc. 29th Conf. Uncertainty Artif. Intell.*, 2013, pp. 262–271.
[27] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1852, Oct. 2007.
[28] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.
[29] X. Li, Y.-Y. Wang, and A. Acero, "Extracting structured information from user queries with semi-supervised conditional random fields," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2009, pp. 572–579.
[30] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "Learning partially-observed hidden conditional random fields for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 533–540.
[31] J. Nicolle, K. Bailly, and M. Chetouani, "Real-time facial action unit intensity prediction with regularized metric learning," *Image Vis. Comput.*, vol. 52, pp. 1–14, Aug. 2016.
[32] S. Eleftheriadis, O. O. Rudovic, M. P. Deisenroth, and M. Pantic, "Variational Gaussian process auto-encoder for ordinal prediction of facial action units," in *Proc. Asian Conf. Comput. Vis.*, Taipei, Taiwan, Nov. 2016, pp. 154–170.
[33] F. Zhou, F. De la Torre, and J. F. Cohn, "Unsupervised discovery of facial events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2574–2581.
[34] D. M. Tax, E. Hendriks, M. F. Valstar, and M. Pantic, "The detection of concept frames using clustering multi-instance learning," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2917–2920.
[35] R. Zhao, Q. Gan, S. Wang, and Q. Ji, "Facial expression intensity estimation using ordinal information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3466–3474.
[36] K. Sikka, A. Dhall, and M. Bartlett, "Weakly supervised pain localization using multiple instance learning," in *Proc. IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.*, Apr. 2013, pp. 1–8.
[37] R. Winkelmann and S. Boes, *Analysis of Microdata*. Berlin, Germany: Springer, 2006.
[38] R. Gupta, A. A. Diwan, and S. Sarawagi, "Efficient inference with cardinality-based clique potentials," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 329–336.
[39] R. H. Byrd, J. Nocedal, and R. B. Schnabel, "Representations of quasi-newton matrices and their use in limited memory methods," *Math. Program.*, vol. 63, nos. 1–3, pp. 129–156. 1994.
[40] D. Tarlow, K. Swersky, R. S. Zemel, R. P. Adams, and B. J. Frey, "Fast exact inference for recursive cardinality models," in *Proc. Conf. Uncertainty Artif. Intell.*, 2012, pp. 825–834.
[41] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Proc. IEEE Int. Conf. Workshops Automat. Face. Gesture Recognit.*, 2015, pp. 1–8.
[42] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.
[43] M. Kim and V. Pavlovic, "Structured output ordinal regression for dynamic facial emotion intensity prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 649–662.
[44] K. M. Prkachin, "The consistency of facial expressions of pain: A comparison across modalities," *Pain*, vol. 51, no. 3, pp. 297–306, 1992.
[45] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.
[46] Z. Tosér, L. A. Jeni, A. Lőrincz, and J. F. Cohn, "Deep learning for facial action unit detection under large head poses," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 359–371.
[47] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun./Jul. 2016, pp. 1535–1543.
[48] S. Han, Z. Meng, A.-S. Khan, and Y. Tong, "Incremental boosting convolutional neural network for facial action unit recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 109–117.

Adria Ruiz received the Ph.D. degree in information and communication technologies from Universitat Pompeu Fabra, Barcelona, in 2017. He is currently a Post-Doctoral Researcher with the THOTH Team, INRIA, Grenoble. His research interests include general machine learning topics and their application to computer vision problems.

Ognjen (Oggi) Rudovic received the Ph.D. degree in computing from Imperial College London, U.K., in 2014. He is currently a Marie Curie Post-Doctoral Fellow with the MIT Media Laboratory, Affective Computing Group. His research interests are in machine learning and computer vision, and their applications to human–robot interaction, health-care, and personalized learning.

Xavier Binefa received the Ph.D. degree in computer vision from the Universitat Autonoma de Barcelona in 1996. He was an Associate Professor with the Computer Science Department until 2009, when he was contracted by Universitat Pompeu Fabra as an Associate Professor at the Information and Communication Technologies Department, where he leads the Cognitive Media Technologies Group.

Maja Pantic is a Professor in affective and behavioral computing at the Department of Computing, Imperial College London, U.K., and at the Department of Computer Science, University of Twente, The Netherlands. She received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She currently serves as the Editor-in-Chief of *Image and Vision Computing Journal*, and as an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS Part B and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.