Fusion of Valence and Arousal Annotations through Dynamic Subjective Ordinal Modelling

Adria Ruiz, Oriol Martinez, Xavier Binefa and Federico M. Sukno
Department of Information and Communications Technologies. Universitat Pompeu Fabra, Spain

***Abstract*— An essential issue when training and validating computer vision systems for affect analysis is how to obtain reliable ground-truth labels from a pool of subjective annotations. In this paper, we address this problem when labels are given in an ordinal scale and annotated items are structured as temporal sequences. This problem is of special importance in affective computing, where collected data is typically formed by videos of human interactions annotated according to the Valence and Arousal (V-A) dimensions. Moreover, recent works have shown that inter-observer agreement of V-A annotations can be considerably improved if these are given in a discrete ordinal scale. In this context, we propose a novel framework which explicitly introduces ordinal constraints to model the subjective perception of annotators. We also incorporate dynamic information to take into account temporal correlations between ground-truth labels. In our experiments over synthetic and real data with V-A annotations, we show that the proposed method outperforms alternative approaches which do not take into account either the ordinal structure of labels or their temporal correlation.**

I. INTRODUCTION

The analysis of human non-verbal behavior is a topic of increasing importance in computer vision. Concretely, facial expressions and body gestures are accepted to carry important information about human emotions [1] and, thus, its automatic understanding has a wide range of potential applications in the affective computing field. In this context, a lot of research has focused on building computer vision systems able to map non-verbal behavior to a representation of human affect [2]. One of the most popular representations for this purpose is the Valence-Arousal (V-A) space [3].

Formalized by Russell through the Circumplex model of affect, Valence and Arousal have been identified as the underlying dimensions of human emotion. Valence refers to how pleasant or unpleasant is an affective state while Arousal indicates the activation or deactivation level. These two dimensions have been consistently identified in experiments across various modalities [4]–[6], which supports their validity. On the downside, Valence and Arousal are abstract dimensions whose exact meaning, apart from subjective, is not common knowledge (e.g. as opposed to well known emotions such as happiness, fear, etc).

A. Motivation

An essential issue when training and validating computer vision systems based on the V-A representation, is how to obtain ground-truth annotations from collected data (e.g, videos of human interactions). This is typically addressed based on manual annotations from expert human observers.

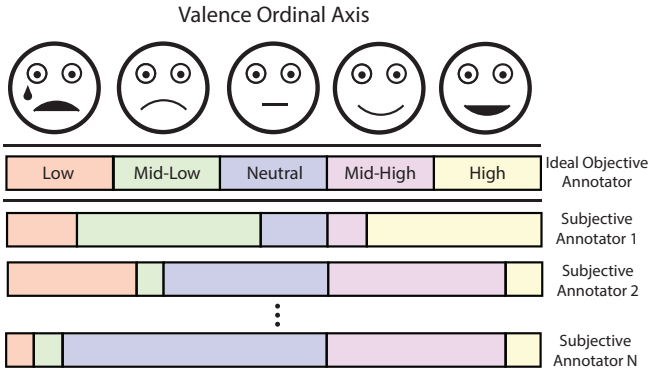

Fig. 1. Illustration of the ordinal subjective assumption to fuse Valence and Arousal annotations. While the *objective* distance between consecutive ordinal labels is hypothetically uniform, each observer has his/her own perception of both the position and extent of them. As stated in [7], there is no justification for the assumption that subjective annotations follow a linear scale (e.g. the perceived distance between pleasant and neutral not necessarily matches the one between neutral and unpleasant). Thus, the only assumption we make in the proposed model is that the order of perceived labels is maintained across annotators, not their distances.

However, labels obtained in this way are evidently subjective and have been shown to suffer from large inter-observer variations [8]–[11].

Subjectivity is unavoidable as it is inherent to affect annotations, regardless of using V-A or any other representation. However, it has been shown that the consistency of subjective annotations can be considerably improved if these are performed based on discrete (instead of continuous) labels and maintaining their ordinal relations [11], [12]. The reasons behind this finding seem related to the invalid assumptions underlying the use of both continuous and non-ordinal (nominal) labels. For instance, nominal labels assume the same degree of confusion between neighbouring and far away labels, while continuous methods assume a linear relation between the true labels and the subjective perception of annotators [13].

Even if following the above recommendations, it is not possible to rely on individual annotations to obtain reliable ground truth. Therefore, consensus from pools of observers are usually preferred. Nevertheless, such consensus is not straightforward to obtain and the problem of fusing annotations from multiple observers has attracted considerable attention [14]–[26]. However, a vast majority of algorithms treat annotated labels simply as nominal classes, e.g. without taking into account their ordinal relations. As explained above, this has been found suboptimal for affective annota-

tions, both theoretically [7], [13] and empirically [11], [27].

B. Contributions

In this paper, we present a novel probabilistic framework to address the fusion of V-A annotations from multiple human observers. The main contributions of this work are summarised as follows:

- The proposed method explicitly considers the ordinal structure in V-A labels and models the gap between the labeling scale and the subjective perception of each annotator. Fig. I illustrates this concept. While the fusion of ordinal labels has been investigated in other domains [22], to the best of our knowledge, this is the first time it is applied to V-A annotations.
- In contrast to previous methods in annotation fusion, the presented framework is able to exploit dynamic information present in temporal sequences of annotations. Despite this information is irrelevant in other applications, it is of special importance in the context of V-A label fusion, where data typically consist of annotated videos of human interactions.
- In our experiments over synthetic and real data annotated with ordinal V-A labels, we show the superior performance of the presented method with respect to alternative approaches that ignore either the ordinal structure of labels or the dynamic information.

II. RELATED WORK

A. Valence and Arousal annotations

The use of dimensional approaches to represent emotions has increasingly gained popularity in the context of affective computing and related fields [28]. Several possible dimensions have been proposed to represent affect, with Valence and Arousal emerging as the most popular ones [10], [29]–[31]. However, in spite of their widespread use, generating reliable annotations in V-A space has proven challenging, not only in terms of achieving consensus but even in the way to address data annotation.

Based on the original definition by Russell [3], several authors directly targeted annotations in V-A space in the continuous domain [29], [30], [32]. This means that each (human) observer is asked to rate a video (either in segments or in continuous time) with points in $\mathbb{R}^2$, typically ranging between -1 and 1 in each axis and constrained to be within the unit circle [33]. Such annotations have proven very variable, producing large dispersions between annotators, even when trying to map a single specific emotion into V-A space [34]. Some methods for the fusion of multiple continuous annotations have been proposed [9], [35], but their ability to model annotator's subjectivity is limited. Indeed, while fusion in continuous space might seem attractive from an ordinal perspective, it actually implies assuming a linear relationship between the ground truth and the observer's annotation. This is implicit in the work by Nicolaou et al. [9] who, moreover, provides only an estimate of how annotations change over time but lose their actual scale, which makes it impossible to apply their method to obtain a properly scaled consensus. In a very recent work, Gupta et al. [35] make the linear mapping between observations and ground truth explicit. Linearity is indeed a key assumption to derive their equations within an EM framework. Extending the mapping to consider non-linear relations would allow a better modeling of annotator's subjectivity, but this possibility has so far not been explored.

On the other hand, some alternatives to the continuous representation of V-A have been recently explored. For example, Baveye et al. [31] provide a ranking between events based on pair-wise comparison from a large number of observers. Thus, they do not provide actual labels but an ordering of events in V-A axes. This strategy is motivated by the argument that human observers are better at producing relative (e.g. pair-wise) comparisons than absolute ratings. Yannakakis et al. [11] provide experimental evidence to support this theory. They compared the V-A annotations performed in continuous space against annotations produced in a ordinal discretized V-A space and found the latter to be clearly more consistent. Interestingly, they also showed that if annotations are performed in continuous V-A space and then discretized, their consensus also improves but not as much as in the case of annotating directly in the discretized ordinal space. Recent efforts in producing V-A labellings have followed this direction [10].

Given the aforementioned advantages of using an ordinal scale to represent Valence and Arousal dimensions, in this paper we focus on the problem of fusing annotations which are given according to discrete but ordered categories. As will be discussed in Sec. II-B, this problem has been recently applied in different domains. However, as far as we know, this is the first time it is explored in the context of V-A affect annotations.

B. Discrete label fusion

Fusion of discrete annotations is a necessity for several fields and it is especially important when the true labels (*ground truth*) are unknown. In such cases, we wish to estimate the ground truth by merging the estimates (annotations) from a number of observers. The basic intuition is that given a sufficient number of observers, we should be able to extract a consensus from their annotations that is reasonably close to the ground truth. Straight-forward solutions to label fusion include averaging, majority voting and extensions such as weighting [14], [15] or iterative outlier removal [16].

A more principled solution consists on adopting a probabilistic framework that jointly estimates the ground truth and the annotators' subjective perception of labels. The latter is done by means of modelling a conditional probability which indicates, for each observer, what is the likelihood of annotating/perceiving a value given a fixed ground-truth label (*annotator's perception model*). Several approaches, have followed this line, among which STAPLE (Simultaneous truth and performance level estimation) is the most popular [17]. STAPLE employs an expectation maximization (EM) algorithm to iteratively estimate the ground truth and perception model parameters, such that the probability of

the observed annotations is maximized. STAPLE has been successfully used in several applications and numerous extensions to the framework have also been proposed. Among the most notable ones we can cite variants for handling partial observations [19], variations or instabilities [20], [21], [25] in annotator performances and variable difficulty throughout the annotation task [26].

When focusing on ordinal affective annotations, an issue of particular importance is the fact that labels are not simply unrelated categories but, instead, they naturally follow a relative ordering. For example, in the Valence axis, *pleasant* is further from *unpleasant* than from *neutral*. Ignoring the ordinal nature of labels has been found suboptimal, not only within affective computing but for subjective annotations in general [13]. However, this issue has been largely overlooked in the label fusion literature. Among recent efforts to incorporate ordinal constraints we find the methods from Zhou et al. [22], based on entropy optimization within a mini-max framework, Metrikov et al. [23], based on latent trait models and Lakshminarayanan et al. [24], based on Bayesian inference. The Ordinal Min-Max Entropy method [22] incorporates ordinal constraints on the annotators' perception model by means of an auxiliary variable that converts multi-label comparisons into a binary problem and optimize the conditional entropy jointly across all possible binary splittings. More closely related to our approach, [23] and [24] use Gaussian priors to model the probabilistic labeling of each annotator conditional to the true (but unknown) labels. However, strictly speaking, the approach from Metrikov et al. [23] cannot ensure that ordinal constraints hold, since the Gaussian models for each label are completely independent of each other. Lakshminarayanan et al. [24] resolve this by a mapping strategy based on pre-defined thresholds that naturally follow the desired ordering of labels, but this limits the flexibility of the approach to model annotator differences.

In contrast to the previously described works, our method employs an ordered probit model [36] to explicitly incorporate ordinal constraints in the annotators' perception model. This approach is both flexible enough to account for annotator-specific differences in perception while still ensures that ordinality is strictly fulfilled. Moreover, the presented framework is also able to incorporate dynamic information, useful when dealing with temporal sequences of annotations. To the best of our knowledge, temporal modelling has not been considered before in the context of ordinal annotation fusion.

III. Problem definition

Following, we formally describe the annotation fusion problem addressed in this work. We assume that a training set of N annotated sequences $\mathbb{D} = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, ..., \mathbf{D}^{(N)}\}$ is provided. Concretely, each $\mathbf{D^{(i)}}$ is an $A \times T$ matrix containing a set of T annotations for a total of A observers, where T is the number of items per sequence [1]. From now on, we will refer to the the label assigned by annotator a to the item t as $D_{at}^{(i)}$. Moreover, we consider the scenario where $D_{at}^{(i)} \in \{0 \prec ... \prec l \prec L\}$ is an ordinal variable taking L possible values.

Similar to previous works on label fusion [17], we aim to learn a parametric model maximizing the log-likelihood over the training set $\mathbb{D}$ as:

$$\arg\max_{\theta} \; L(\mathbb{D} \mid \theta) = \log(p(\mathbb{D} \mid \theta)) \tag{1}$$

where θ is the set of model parameters. For this purpose, we define for each sequence a new set of latent variables $\mathbf{g^{(i)}} = \{g_1^{(i)}, g_2^{(i)}, ..., g_T^{(i)}\}$ representing the ground-truth ordinal labels for each item. Given $\mathbf{g^{(i)}}$, the log-likelihood can be expressed as:

$$L(\mathbb{D} \mid \theta) = \sum_{i=1}^{N} \log \left(\sum_{\mathbf{g}} \prod_{a} p(\mathbf{D}_a^{(i)} \mid \mathbf{g}; \theta) p(\mathbf{g}) \right) \tag{2}$$

by assuming conditional independence between observer annotations for each sequence and marginalizing over all the possible latent ground-truth labellings $\mathbf{g}$.

Given the parameters θ that maximize Eq. 2, the probability of a given ground-truth labelling $\mathbf{g}^{(i)}$ for a given sequence can be obtained from:

$$p(\mathbf{g}^{(i)} | \mathbf{D}^{(i)}, \theta) = \frac{\prod_a p(\mathbf{D}_a^{(i)} | \mathbf{g}^{(i)}; \theta) p(\mathbf{g}^{(i)})}{\sum_{\mathbf{g}} \prod_a p(\mathbf{D}_a^{(i)} | \mathbf{g}; \theta) p(\mathbf{g})} \tag{3}$$

where $p(\mathbf{g})$ defines a prior over $\mathbf{g}$.

IV. Static Ordinal Annotation Fusion

Following, we describe the proposed Static Ordinal Annotation Fusion (SOAF). This model aims to solve the problem defined in Sec. III by ignoring temporal information. For this purpose, we assume that item ground-truth labels $g_t^{(i)}$ within a sequence are independent. Under such assumption, Eq. 2 can be expressed as:

$$L = \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left(\sum_{l} \prod_{a} p(\mathbf{D}_{at}^{(i)} \mid g_t^{(i)} = l, \theta^a) p(g_t^{(i)} = l) \right), \tag{4}$$

Note that we have defined a set of independent parameters $\theta = \{\theta^1, \theta^2, ..., \theta^A\}$ for each annotator. These parameters model the conditional probability $p(\mathbf{D}_{at} | g_t, \theta^a)$ and, thus, describe the subjective perception of each annotator a for a given latent ground-truth ordinal label (annotator perception model). Under the defined model, the ground-truth probability for any sequence item can be easily computed as:

$$p(g_t^{(i)} = l | \mathbf{D}^{(i)}, \theta) = \frac{\prod_a p(\mathbf{D}_{at}^{(i)} | g_t^{(i)} = l)}{\sum_{l'} \prod_a p(\mathbf{D}_{at}^{(i)} | g_t^{(i)} = l')} \tag{5}$$

by assuming an uniform prior distribution over all $p(g_t^{(i)})$

[1] For notation simplicity, we use the same number of annotators A and T for each sequence. However, the proposed methods can also handle cases where they vary across sequences

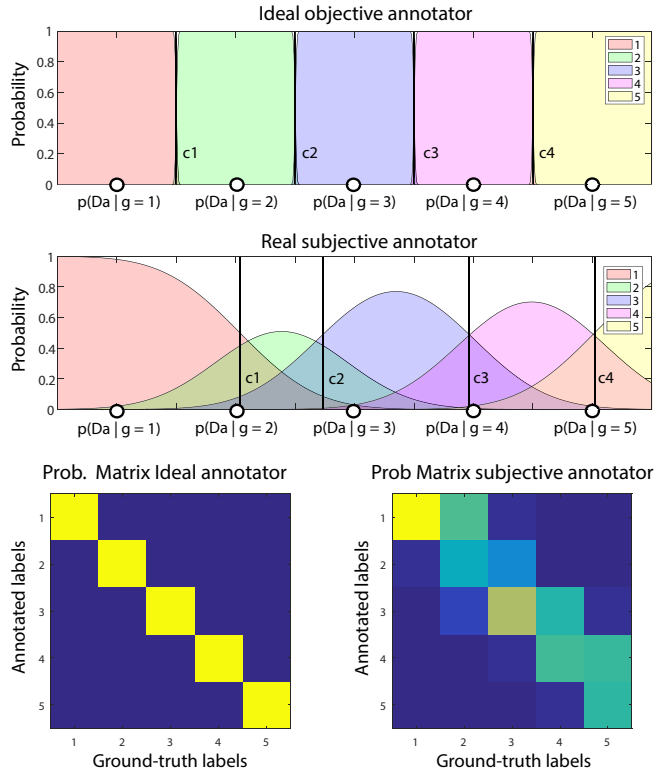

Fig. 2. Illustration of the employed ordered probit model defining the annotator perception models $p(\mathbf{D}_{at}|g_t, \theta^a)$. Top: Ideal objective annotator perceiving equally-distant ordinal labels with no uncertainty ($\sigma = 0$). Middle: Real annotator where the perception of different labels is non-linear but follows ordinal constraints ($\sigma \approx 0.5$ modelling perception noise). Note that for both annotators, the perceived distance between ordinal values are determined by thresholds $\mathbf{c}$. The monotonically increasing constraints over these thresholds ensure that the likelihood of perceived labels are ordered. Bottom: For both cases, matrices representing the conditional probabilities $p(\mathbf{D}_{at}|g_t, \theta^a)$ for each pair of ground-truth and perceived ordinal labels.

A. Ordinal annotator perception model

In order to incorporate the ordinal constraints into the annotator's perception model, we propose to use an ordered probit model [36] to define conditional probabilities $p(\mathbf{D}_{at} = l|g_t = l', \theta^a)$ as:

$$p(\mathbf{D}_{at} = l|g_t = l', \theta^a) = \Phi\left(\frac{c_l^a - l'}{\sigma^a}\right) - \Phi\left(\frac{c_{(l-1)}^a - l'}{\sigma^a}\right). \tag{6}$$

Here, $\Phi(\cdot)$ denotes the normal cumulative distribution function (CDF), and $\theta^a = \{\mathbf{c^a}, \sigma^a\}$ is the set of annotator parameters. Specifically, $\mathbf{c}^a = \{c_0^a = -\infty \leq c_1^a \leq \cdots \leq c_L^a = \infty\}$ are monotonically increasing thresholds dividing a continuous line into L bins corresponding to different ordinal values. The difference between the two CDFs provides the probability of a perceived label l given the ground-truth ordinal value l'. Moreover, $\sigma^a > 0$ is the standard deviation of Gaussian noise modelling uncertainty in the observer annotations (see Fig. 2 for an illustration). This model has been previously explored in the context of facial expression intensity estimation [37].

In order to ensure $\sigma > 0$ and the monotonically increasing constraints of thresholds $\mathbf{c}$, we use a re-parametrization strategy similar to [38]. Concretely, we define $c_l = c_1 + \sum_{s=1}^{l-1} \delta_s^2$ and $\sigma = \tau^2$. With this parametrization, the maximization of Eq. 4 becomes an unconstrained optimization problem which can be solved as described in the following section.

B. Learning

In order to learn the optimal model parameters θ given a set $\mathbb{D}$, we use standard gradient ascent. Specifically, we employ the LBFGS Quasi-Newton method [39], which generally provides a higher-convergence rate than first-order approaches. The derivatives of the log-likelihood function $L(\mathbb{D} \mid \theta)$ w.r.t the annotator parameters θ^a can be expressed as:

$$\frac{\delta L}{\delta \theta^a} = \sum_{i,t,l} \frac{p(g_t^{(i)} = l|\mathbf{D}^{(i)}, \theta)}{p(\mathbf{D}_{at}^{(i)}|g_t^{(i)} = l)} \cdot \frac{\delta p(\mathbf{D}_{at}^{(i)}|g_t^{(i)} = l)}{\delta \theta^a} \tag{7}$$

where the gradients $\frac{\delta p(\mathbf{D}_{at}^{(i)}|g_t^{(i)}=l)}{\delta \theta^a}$ with the defined ordinal probit model can be easily computed as detailed in [38].

V. Dynamic Ordinal Annotation Fusion

One of the main assumptions made in SOAF (Sec IV), is that ground-truth latent variables g_t^i for a given sequence are independent. This assumption is suboptimal in temporal sequences (e.g. videos) where ground-truth labels tend to be temporally-correlated. In order to incorporate dynamic information, we extend SOAF to Dynamic Ordinal Annotation Fusion (DOAF). For this purpose, we follow a first-order Markovian assumption where a given g_t is dependent on the previous ground-truth label g_{t-1}. In DOAF, the log-likelihood can be expressed as:

$$L = \sum_{i=1}^{N} \log\left(\sum_{\mathbf{g}} \left[p(g_1^{(i)}) p(\mathbf{D}_{:1}^{(i)} \mid g_t^{(1)}, \theta^a) \cdot \cdot \prod_{t=2}^{T} p(\mathbf{D}_{:t}^{(i)} \mid g_t^{(i)}, \theta^a) p(g_t^{(i)}|g_{t-1}^{(i)}; \theta^D)\right]\right), \tag{8}$$

where $p(\mathbf{D}_{:t}^{(i)} \mid g_t^{(i)}; \theta^a) = \prod_a p(\mathbf{D}_{at}^{(i)} \mid g_t^{(i)}; \theta^a)$ and $\theta^D = \{\alpha_{1|1}, \ldots, \alpha_{l|l'}, \ldots, \alpha_{L|L}\}$ is a new set of parameters defining the label transition probabilities $p(g_t|g_{t-1})$. Concretely, we use a soft-max function:

$$p(g_t = l|g_{t-1} = l'; \theta^D) = \frac{e^{\alpha_{l|l'}}}{\sum_s e^{\alpha_{s|l'}}}, \tag{9}$$

which ensures the conditional probability constraints $\sum_l p(l|l') = 1$ and $p(l|l') \geq 0$.

Note that DOAF is a special case of a Hidden Markov Model [40] (HMM) where latent states represent the ground-truth labels and emission probabilities are defined by $p(\mathbf{D}_{:t}^{(i)} \mid \theta^a)$. Therefore, marginal probabilities $p(g_t^{(i)}|\mathbf{D}^{(i)}, \theta)$ of item latent labels can be computed using the forward-backward algorithm [41] employed in HMMs. Similar to SOAF, we assume a uniform prior distribution over initial latent ordinal states $p(g_1)$.

A. Learning

In DOAF, we use a similar learning procedure than the one described in Sec. IV-B for SOAF. In this case, the gradients of parameters θ^a can be computed as defined in Eq. 7. However, marginal probabilities $p(g_t^{(i)}|\mathbf{D}^{(i)},\theta)$ need to be obtained using the forward-backward procedure. On the other hand, the gradient of transition parameters θ^D can be computed as:

$$\frac{\delta L}{\delta \alpha_{s|s'}} = \sum_{i,t,l,l'} \Big[\mathbb{1}(s'=l') p_{l,l'}^{it} [\mathbb{1}(s=l) - p_{s|s'}^{it}] \Big], \quad (10)$$

where $\mathbb{1}(\cdot)$ is an indicator function and

$$p_{l|l'}^{it} = p(g_t^{(i)} = l | g_{t-1}^{(i)} = l'; \theta^D) \quad (11)$$

$$p_{s,s'}^{it} = p(g_t^{(i)} = l, g_{t-1}^{(i)} = l' \mid \mathbf{D}^{(i)}; \theta). \quad (12)$$

Again, marginal probabilities $p(g_t^{(i)} = l, g_{t-1}^{(i)} = l' \mid \mathbf{D}^{(i)})$ can be computed with the forward-backward procedure.

VI. EXPERIMENTS

A. Evaluation criteria and metrics

In order to compare the proposed SOAF and DOAF frameworks with alternative approaches, we use two different criteria. In the first one, we evaluate the prediction of ground-truth labels. Formally, given the learned annotator perception models represented by $p(\mathbf{D}_{at}|g_t)$, we estimate the most likely ground-truth labels $\mathbf{g}_{pred}^{(*)}$ for a new test sequence $\mathbf{D}^{(*)}$ (see Eq. 3). Assuming that we know the real annotations $\mathbf{g}_{real}^{(*)}$ for $\mathbf{D}^{(*)}$, we compare predictions and real labels for each sequence item. For that purpose, we employ standard metrics used in the context of ordinal regression [42], [43]. Concretely, we use the Pearson's Correlation Coefficient (CORR), Mean Absolute Error (MAE) and the Intra-Class-Correlation Coefficient (ICC).

Despite the main goal of annotation fusion methods is to predict ground-truth labels, evaluation under the aforementioned criterion is generally not feasible in real scenarios, since the actual $\mathbf{g}_{real}^{(*)}$ is not known. In order to compare different methods in the context of V-A annotations, we use an alternative evaluation criterion. It is based on the assumption that, given the perception model $p(\mathbf{D}_{at}|g_t)$, we should be able to predict new annotations for the observer a given a predicted ground truth $\mathbf{g}_{pred}^{(*)}$. Formally, the predicted annotation for a given sequence and observer can be computed as:

$$p(\mathbf{D}_a^{(*)}|\mathbf{D}_{\forall a' \neq a}^{(*)}) = \sum_{\mathbf{g}^{(*)}} p(\mathbf{D}_a^{(*)}|\mathbf{g}^{(*)}) p(\mathbf{g}^{(*)}|\mathbf{D}_{\forall a' \neq a}^{(*)}) \quad (13)$$

where $\mathbf{D}_{\forall a' \neq a}^{(*)}$ refers to the sequence annotations for all the observers except a. Given that we know $\mathbf{D}_a^{(*)}$ for a test sequence, the same metrics previously described can be used to evaluate model's performance. Note that this criteria jointly evaluates the annotators model and ground truth estimation, since both are needed to estimate the annotator labellings. For instance, even if we had the optimal perception model for a given annotator, it would be impossible to correctly generate his labelling for a given test sequence if the estimated ground-truth was not accurate.

B. Baselines

In our experiments, we compare the proposed SOAF and DOAF methods with alternative approaches that ignore either the ordinal structure of labels (nominal) or the dynamic information (static). Following, we describe them.

Majority Voting (MV): The ground-truth labelling is predicted with a majority voting strategy. Concretely, the estimated label for a given item is chosen as the majority ordinal level across all the annotators. Given that this approach does not explicitly compute the annotator's perception model $p(\mathbf{D}_{at}|g_t)$, we empirically compute it from the training annotated sequences and the estimated ground-truth. MV follows an static-nominal assumption.

STAPLE: This method is one of the most popular approaches for fusing annotations with nominal labels. (see Sec. II-B and [17]). We have used our own implementation of this method which also follows a static-nominal assumption.

Static Nominal Annotation Fusion (SNAF): This approach is equivalent to the proposed SOAF model but modelling ordinal labels as nominal. Concretely, $p(\mathbf{D}_{at}|g_t)$ for each annotator is defined using a parametrized soft-max function (see Eq. 9). It can be easily shown that SNAF maximizes the same log-likelihood function as STAPLE. However, SNAF is trained using gradient-ascent whereas STAPLE uses an Expectation-Maximization algorithm.

Dynamic Nominal Annotation Fusion (DNAF): In this case, DNAF is equivalent to the proposed DOAF model but modelling ordinal labels as nominal similar to SNAF. Therefore, we can consider DNAF a dynamic-nominal approach.

Ordinal Minimax Conditional Entropy (OMME): This method can be considered the state-of-the-art approach for static ordinal annotation fusion (see Sec. II-B). In our experiments, we use the implementation provided by the authors of the original paper. Similarly to the case of MV, we empirically compute the annotator conditional probabilities $p(\mathbf{D}_{at}|g_t)$.

C. Synthetic Experiments

To validate the benefits of the proposed framework while fusing ordinal annotations of temporal sequences, we have performed a set of experiments using synthetic data. The use of these data allows to evaluate the performance of different approaches while predicting latent ground-truth labels. As explained in Sec. VI-A, this ground-truth is not known in real data and, therefore, it is not feasible to evaluate methods according to this criterion.

Data generation and experimental setup: In order to create a synthetic dataset of annotated sequences $\mathbb{D}$, we use the following procedure. Firstly, we generate a DOAF model (see Section V) by randomly defining a set of parameters $\theta = \{\theta^1, ..., \theta^A, \theta^D\}$. The number of ordinal levels and annotators has been set to $L = 6$ and $A = 4$ respectively.

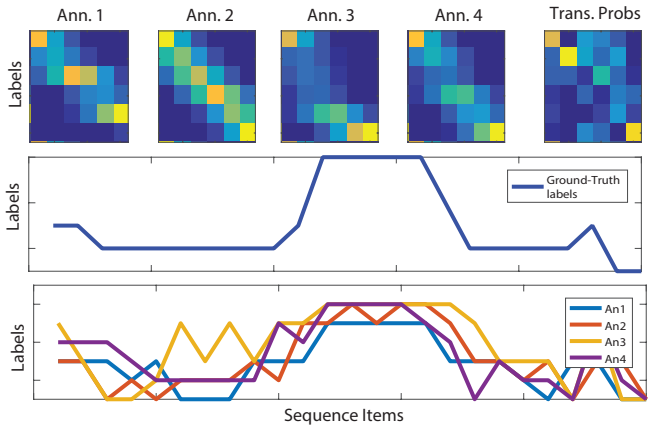

Fig. 3. Illustration of the process followed to generate synthetic data sequences. From top to bottom: (i) Matrices representing the annotator perception models (A=4) and temporal transition probabilities from a randomly generated DOAF model. (ii) Example of a ground-truth sequence sampled according to the defined transition probabilities. (iii) Randomly generated annotations according to the defined ground-truth sequence and perception models.

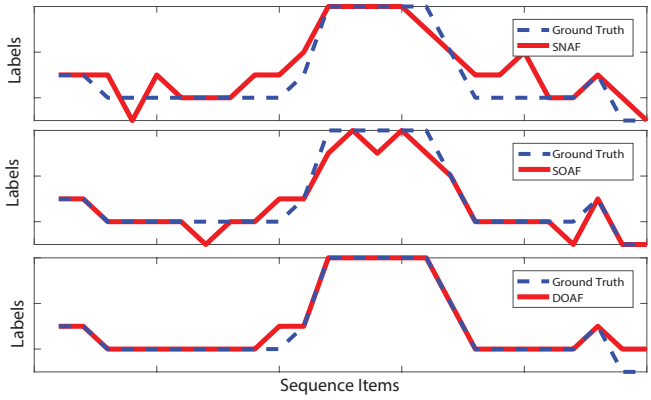

Fig. 4. Examples of ground-truth predictions in a synthetic sequence for SNAF, SOAF and DOAF. Note that SOAF predicts more accurately the actual latent ordinal levels than SNAF, which models labels as nominal variables. Moreover, SOAF predictions tend to be less temporally smooth than in the DOAF case. This is because the latter incorporates dynamic information which takes into account the conditional dependencies between temporally consecutive items in the sequence.

Secondly, for each sequence $\mathbf{D}^{(\mathbf{i})}$ we sample the ground-truth labels $\mathbf{g}_{real}^{(*)}$ by using the conditional probabilities $p(g_t|g_{t-1};\theta^D)$ and a uniform prior distribution over $p(g_1)$. Sequences length has been set to $T = 50$. Finally, for each observer we generate his annotations $\mathbf{D}_{\mathbf{a}}^{(\mathbf{i})}$ by sampling from his perception model $p(\mathbf{D}_{at}|g_t)$ and $\mathbf{g}_{real}^{(*)}$. Figure 3 illustrates this process. Using this procedure, we have generated 100 synthetic datasets with 10 and 20 sequences for training and testing respectively (randomly generating different DOAF parameters for each dataset). Training sequences are used to learn the different model parameters whereas test sequences are used to compute the different metrics described in VI-A

Results and discussion: Table I shows the average results over the 100 synthetic datasets . By looking into the results of the compared methods, we can derive the following conclusions. Firstly, MV obtains the worst performance among all metrics. This was expected since it follows a static-nominal approach and does not take into account the perception model for each annotator. Secondly, static-nominal approaches such as SNAF and STAPLE generally perform worse than static-ordinal methods (OMME and SOAF). This shows the importance of taking into account the ordinal structure of labels in this kind of problems. Thirdly, note that the dynamic-nominal approach DNAF obtains slightly better results than all the other static-nominal approaches by considering the dynamic information present in temporal sequences of annotations. Finally, DOAF obtains the best performance in all cases by taking into account both dynamic information and introducing ordinal constraints in the annotator perception models. To illustrate this conclusions, we show in Fig. 4 an example of qualitative results obtained by SNAF, SOAF and DOAF in the synthetic test sequence shown in Fig. 3.

TABLE I

OBTAINED RESULTS RESULTS ON SYNTHETIC GENERATED DATA

	Ground-truth Predictions			Annotation Predictions		
	CORR	**MAE**	**ICC**	**CORR**	**MAE**	**ICC**
MV	0.859	0.648	0.853	0.674	0.944	0.673
STAPLE [17]	0.869	0.648	0.862	0.737	0.790	0.736
SNAF	0.867	0.653	0.860	0.737	0.792	0.736
DNAF	0.901	0.501	0.895	0.740	0.783	0.739
OMME [22]	0.901	0.516	0.894	0.738	0.790	0.737
SOAF	0.928	0.420	0.923	0.744	0.771	0.743
DOAF	**0.941**	**0.259**	**0.940**	**0.756**	**0.746**	**0.755**

D. Valence and Arousal annotations fusion

In order to evaluate the proposed method in the context of V-A annotation fusion, we have used the database described in [10]. To our knowledge, this is the largest database providing a set of annotated videos using ordinal V-A labels. Moreover, we have discarded databases with continuous ratings, such as SEMAINE [30], because quantization of continuous annotations does not result in ordinal data. In contrast, ratings from [10] were annotated according to a small set of ordered discrete labels, which can be validated as an ordinal setting [11]. The rationale is as follows: with only a few labels to choose from, you intuitively compares among them. As the choices increase, this task is more difficult. In the extreme case (continuous), it is impossible for an annotator to keep strict ordinality and subsequent discretization cannot fix that.

Database and experimental setup: The database consists of 64 videos of human interactions -with a total duration of approximately 3.5h- annotated by a maximum of 11 human experts. Valence and Arousal dimensions are labelled on different axes and represented with a set of 7 ordinal labels: {positive, half positive, mild positive, neutral, mild negative half negative, negative}. We have performed an 8-fold cross-validation for both dimensions, where 56 videos have been used for training and 8 for testing. Similar to synthetic experiments, training videos are used to learn the annotators' perception models which are then employed for evaluation on test sequences. In the case of the Arousal dimension, we do not use the lowest negative label since it never appears in

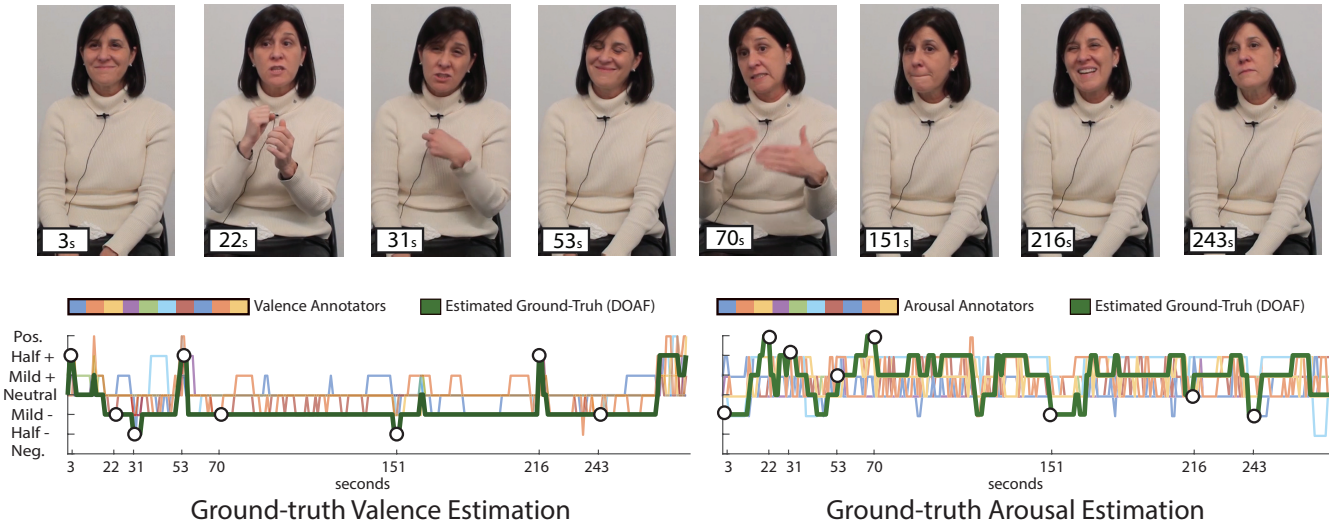

Fig. 5. Estimated ground-truth from a set of V-A annotations in a test video. Despite the noisy subjective annotations provided by different observers, our method is able to estimate a sequence of ground-truth labels coherent with the non-verbal behavior displayed by the subject.

the dataset. In order to reduce computational complexity and remove temporal redundancy, all sequences have been subsampled to only contain time instances where any annotator reported a change in the affective state.

Results and discussion: Given that ground-truth labels $\mathbf{g}_{real}^{(*)}$ for real sequences are unknown, we use the second evaluation criterion described in Sec. VI-A. Table II shows the results obtained for each affective dimension following the previously described cross-validation procedure. We can see that the static-ordinal approaches OMME and SOAF generally obtain better performance than static-nominal methods (SNAF and STAPLE). Indeed, between these four methods, the best performance for any metric is always obtained by either OMME or SOAF. This shows the advantages of considering labels' ordinal structure in this context. Secondly, DOAF outperforms static-ordinal approaches by incorporating dynamic information into the ground-truth labelling estimation. However, note that DNAF does not actually outperforms SNAF. This suggest that the advantage of modelling temporal correlations can only be fully achieved if appropriately considering labels ordinality. In conclusion, our results show the benefits of the proposed DOAF model for V-A annotations fusion. Fig. 5 shows an illustrative example of the estimated ground-truth labels by DOAF in a test sequence.

TABLE II

OBTAINED RESULTS ON AROUSAL AND VALENCE ANNOTATIONS OF HUMAN INTERACTION RECORDINGS

	Arousal Annotations			Valence Annotations		
	CORR	**MAE**	**ICC**	**CORR**	**MAE**	**ICC**
MV	0.308	0.529	0.300	0.483	0.486	0.471
STAPLE [17]	0.343	0.513	0.337	0.496	0.463	0.479
SNAF	0.359	0.506	0.352	0.493	0.462	0.471
DNAF	0.332	0.538	0.330	0.503	0.482	0.497
OMME [22]	0.352	0.516	0.349	0.514	0.481	0.511
SOAF	0.368	0.497	0.354	0.509	0.454	0.457
DOAF	**0.400**	**0.492**	**0.391**	**0.542**	**0.445**	**0.516**

VII. CONCLUSIONS

In this paper, we propose a novel framework for the fusion of ordinal annotations in temporal sequences. This problem is of special importance in the context of affective computing, where collected data is typically formed by videos of human interactions annotated in terms of V-A affective labels. Recent works have shown that the consistency of V-A annotations can be considerably improved if these are performed based on an ordinal scale. Thus, in contrast to previous methods for annotation fusion, our approach explicitly introduces ordinal constraints into the annotators' perception model and incorporates dynamic information useful when dealing with temporal sequences. In our experiments over synthetic and real data, we show that the proposed method outperforms alternative approaches which do not take into account either the ordinal structure of labels or the dynamic information. Future datasets may benefit from the presented framework as it would help to provide more reliable ground-truth to train and validate automatic affect analysis models.

ACKNOWLEDGMENTS

This work is partly supported by the Spanish Ministry of Economy and Competitiveness under the Ramon y Cajal fellowships and the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), and the Kristina project funded by the European Union Horizon 2020 research and innovation programme under grant agreement No 645012. Adria Ruiz would also like to acknowledge Spanish Government to provide support under grant FPU13/01740.

REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, pp. 92–105, 2011.

[3] J. A. Russell, "A circumplex model of affect.," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, Dec. 1980.

[4] N. Cliff and F. W. Young, "On the relation between undimensional judgments and multidimensional scaling," *Organizational Behavior and Human Performance*, vol. 3, no. 3, pp. 269–285, 1968.

[5] R. S. Green and N. Cliff, "Multidimensional comparisons of structures of vocally and facially expressed emotion," *Perception & Psychophysics*, vol. 17, no. 5, pp. 429–438, 1975.

[6] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-cultural universals of affective meaning*. University of Illinois Press, 1975.

[7] S. Jamieson *et al.*, "Likert scales: how to (ab) use them," *Medical education*, vol. 38, no. 12, pp. 1217–1218, 2004.

[8] L. Devillers, R. Cowie, J. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie, "Real life emotions in french and english tv video clips: an integrated annotation protocol combining continuous and discrete approaches," in *5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, p. 22, 2006.

[9] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014.

[10] F. Sukno, M. Domínguez, A. Ruiz, D. Schiller, F. Lingenfelser, L. Pragst, E. Kamateri, and S. Vrochidis, "A multimodal annotation schema for non-verbal affective analysis in the health-care domain," in *Proceedings of the 1st International Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction*, pp. 9–14, ACM, 2016.

[11] G. N. Yannakakis and H. P. Martínez, "Grounding truth via ordinal annotation," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pp. 574–580, IEEE, 2015.

[12] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–8, IEEE, 2013.

[13] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!," *Frontiers in ICT*, vol. 2, p. 13, 2015.

[14] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solórzano, "Combination strategies in multi-atlas image segmentation: application to brain mr data," *IEEE transactions on medical imaging*, vol. 28, no. 8, pp. 1266–1277, 2009.

[15] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 3, pp. 611–623, 2013.

[16] T. R. Langerak, U. A. van der Heide, A. N. Kotte, M. A. Viergever, M. van Vulpen, and J. P. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple)," *IEEE Transactions on Medical Imaging*, vol. 29, no. 12, pp. 2000–2008, 2010.

[17] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE transactions on medical imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[18] S. K. Warfield, K. H. Zou, and W. M. Wells, "Validation of image segmentation by estimating rater bias and variance," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1874, pp. 2361–2375, 2008.

[19] B. A. Landman, A. J. Asman, A. G. Scoggins, J. A. Bogovic, F. Xing, and J. L. Prince, "Robust statistical fusion of image labels," *IEEE transactions on medical imaging*, vol. 31, no. 2, pp. 512–522, 2012.

[20] A. J. Asman and B. A. Landman, "Formulating spatially varying performance in the statistical fusion framework," *IEEE transactions on medical imaging*, vol. 31, no. 6, pp. 1326–1336, 2012.

[21] O. Commowick and S. K. Warfield, "Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori staple," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 25–32, Springer, 2010.

[22] D. Zhou, Q. Liu, J. C. Platt, and C. Meek, "Aggregating ordinal labels from crowds by minimax conditional entropy.," in *ICML*, pp. 262–270, 2014.

[23] P. Metrikov, V. Pavlu, and J. A. Aslam, "Aggregation of crowdsourced ordinal assessments and integration with learning to rank: A latent trait model," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1391–1400, ACM, 2015.

[24] B. Lakshminarayanan and Y. W. Teh, "Inferring ground truth from multi-annotator ordinal data: a probabilistic approach," *arXiv preprint arXiv:1305.0015*, 2013.

[25] O. Commowick, A. Akhondi-Asl, and S. K. Warfield, "Estimating a reference standard segmentation with spatially varying performance parameters: local map staple," *IEEE transactions on medical imaging*, vol. 31, no. 8, pp. 1593–1606, 2012.

[26] A. J. Asman and B. A. Landman, "Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (collate)," *IEEE Transactions on Medical Imaging*, vol. 30, no. 10, pp. 1779–1794, 2011.

[27] G. N. Yannakakis and J. Hallam, "Ranking vs. preference: a comparative study of self-reporting," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 437–446, Springer, 2011.

[28] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[29] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[30] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[31] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[32] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–8, IEEE, 2013.

[33] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[34] P. Robinson and T. Baltruaitis, "Empirical analysis of continuous affect," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pp. 288–294, Sept 2015.

[35] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations based on ground truth inference and distortion," *IEEE Transactions on Affective Computing*, in press, 2017.

[36] A. Agresti, *Analysis of ordinal categorical data*, vol. 656. John Wiley & Sons, 2010.

[37] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *Proc. Computer Vision and Pattern Recognition*, pp. 2634–2641, IEEE, 2012.

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

[39] R. H. Byrd, J. Nocedal, and R. B. Schnabel, "Representations of quasi-newton matrices and their use in limited memory methods," *Mathematical Programming*, 1994.

[40] L. R. Rabiner and B.-H. Juang, "An introduction to hidden markov models," *ASSP Magazine*, 1986.

[41] D. Barber, *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[42] A. Ruiz, O. Rudovic, X. Binefa, and M. Pantic, "Multi-instance dynamic ordinal random fields for weakly-supervised pain intensity estimation," *Asian Conference On Computer Vision*, 2016.

[43] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 944–958, 2015.