Hidden Conditional Ordinal Random Fields for Sequence Classification

Minyoung Kim and Vladimir Pavlovic

Rutgers University, Piscataway, NJ 08854, USA
{mikim,vladimir}@cs.rutgers.edu
http://seqam.rutgers.edu

Abstract. Conditional Random Fields and Hidden Conditional Random Fields are a staple of many sequence tagging and classification frameworks. An underlying assumption in those models is that the state sequences (tags), observed or latent, take their values from a set of nominal categories. These nominal categories typically indicate tag classes (e.g., part-of-speech tags) or clusters of similar measurements. However, in some sequence modeling settings it is more reasonable to assume that the tags indicate ordinal categories or ranks. Dynamic envelopes of sequences such as emotions or movements often exhibit intensities growing from neutral, through raising, to peak values. In this work we propose a new model family, Hidden Conditional Ordinal Random Fields (H-CORFs), that explicitly models sequence dynamics as the dynamics of ordinal categories. We formulate those models as generalizations of ordinal regressions to structured (here sequence) settings. We show how classification of entire sequences can be formulated as an instance of learning and inference in H-CORFs. In modeling the ordinal-scale latent variables, we incorporate recent binning-based strategy used for static ranking approaches, which leads to a log-nonlinear model that can be optimized by efficient quasi-Newton or stochastic gradient type searches. We demonstrate improved prediction performance achieved by the proposed models in real video classification problems.

1 Introduction

In this paper we tackle the problem of time-series sequence classification, a task of assigning an entire measurement sequence a label from a finite set of categories. We are particularly interested in classifying videos of real human/animal activities, for example, facial expressions. In analyzing such video sequences, it is often observed that the sequences in nature undergo different phases or intensities of the displayed artifact. For example, facial emotion signals typically follow envelope-like shapes in time: `neutral`, `increase`, `peak`, and `decrease`, beginning with low intensity, reaching a maximum, then tapering off. (See Fig. 1 for the intensity envelope visually marked for an facial emotion video.) Modeling such an envelop is important for faithful representation of motion sequences and consequently for their accurate classification. A key challenge, however, is

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part II, LNAI 6322, pp. 51–65, 2010.

that even though the action intensity follows the same qualitative envelope the rates of increase and decrease differ substantially across subjects (e.g., different subjects express the same emotion with substantially different intensities).

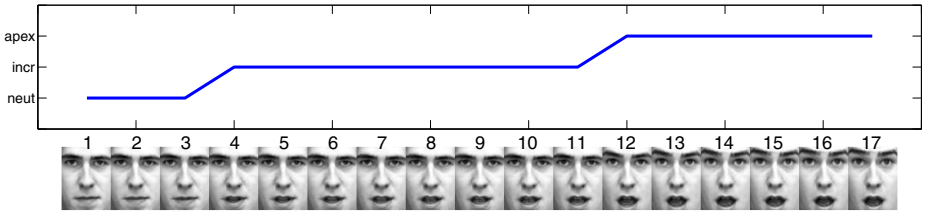


Fig. 1. An example of facial emotion video and corresponding intensity labels. The ordinal-scale labels over time form an intensity envelope (the first half shown here).

We propose a new modeling framework of Hidden Conditional Ordinal Random Fields (H-CORFs) to accomplish the task of sequence classification while imposing the qualitative intensity envelope constraint. H-CORF extends the framework of Hidden Conditional Random Fields (H-CRFs) [12,5] by replacing the hidden layer of H-CRFs category indicator variables with a layer of variables that represent the qualitative but latent intensity envelope. To model this envelope qualitatively yet accurately we require that the state space of each variable be *ordinal*, corresponding to the intensity rank of the modeled activity at any particular time. As a consequence, the hidden layer of H-CORF is a sequence of ordinal values whose differences model qualitative intensity dissimilarities between various stages of an activity. This is distinct from the way the latent dynamics are modeled in traditional H-CRFs, where states represent different categories without imposing their relative ordering. Modeling the dynamic envelope in a qualitative, ordinal manner is also critical for increased robustness. While the envelope could plausibly be modeled as a sequence of real-valued absolute intensity states, such models would inevitably introduce undesired dependencies. In such cases the differences in absolute intensities could be strongly tied to a subject or a manner in which the action is produced, making the models unnecessarily specific while obscuring the sought-after identity of the action.

To model the qualitative shape of the intensity envelope within H-CORF we extend the framework of ordinal regression to structured ordinal sequence spaces. The ordinal regression, often called the preference learning or ranking [6], has found applications in several traditional ranking problems, such as image classification and collaborative filtering [14,2], or image retrieval [7,8]. In the static setting, the goal is to predict the label of an item represented by feature vector $\mathbf{x} \in \mathbb{R}^p$ where the output label bears particular meaning of preference or order (e.g., low, medium or high). The ordinal regression is fundamentally different from the standard regression in that the actual absolute difference of output values is nearly meaningless, but only their relative order matters (e.g., low $<$ medium $<$ high). The ordinal regression problems may not be optimally handled by the standard multi-class classification either because of classifier's ignorance

of the ordinal scale and symmetric treatment of different output categories (e.g., low would be equally different from high as it would be from medium).

Despite their success in static settings (i.e., a vectorial input associated with a singleton output label), ranking problems are rarely explored in structured problems, such as the segmentation of emotion signals into regions of neutral, increasing or peak emotion or actions into different intensity stages. In this case the ranks or ordinal labels at different time instances should vary smoothly, with temporally proximal instances likely to have similar ranks. For this purpose we propose an intuitive but principled Conditional Ordinal Random Field (CORF) model that can faithfully represent multiple ranking variables correlated in a combinatorial structure. The binning-based modeling strategy adopted by recent static ranking approaches (see (2) in Sec. 2.1) is incorporated into our structured models, CORF and H-CORF, through graph-based potential functions. While this formulation leads to a family of log-nonlinear models, we show that the models can still be estimated with high accuracy using general gradient-based search approaches.

We formally setup the problem and introduce basic notation below. We then propose a model for prediction of ordinal intensity envelopes in Sec. 2. Our classification model based on the ordinal modeling of the latent envelope is described in Sec. 3. In Sec. 4, the superior prediction performance of the proposed structured ranking model to the regular H-CRF model is demonstrated on two problems/datasets: emotion recognition from the CMU facial expression dataset [11] and behavior recognition from the UCSD mouse dataset [4].

1.1 Problem Setup and Notations

We consider a K-class classification problem, where we let $y \in \{1, ..., K\}$ be the class variable and $\mathbf{x}$ be the input covariate for predicting y. In the structured problems we assume that $\mathbf{x}$ is composed of individual input vectors $\mathbf{x}_r$ measured at the temporal and/or spatial positions r (i.e., $\mathbf{x} = \{\mathbf{x}_r\}$). Although our framework can be applied to arbitrary combinatorial structures for $\mathbf{x}$, in this paper we focus on the sequence data, written as $\mathbf{x} = \mathbf{x}_1 \dots \mathbf{x}_T$ where the sequence length T can vary from instance to instance. Throughout the paper, we assume a supervised setting: we are given a training set of n data pairs $\mathcal{D} = \{(y^i, \mathbf{x}^i)\}_{i=1}^n$, which are i.i.d. samples from an underlying but unknown distribution.

2 Structured Ordinal Modeling of Dynamical Envelope

In this section we develop the model which can be used to infer ordinal dynamical envelope from sequences of measurement. The model is reminiscent of a classical CRF model, where its graphical representation corresponds to the upper two layers in Fig. 2 with the variables $\mathbf{h} = h_1, \dots, h_T$ treated as observed outputs. But unlike the CRF it restricts the envelope (i.e., sequence of tags) to reside in a space of ordinal sequences. This requirement will impose ordinal, rank-like, similarities between different states instead of the nominal differences of

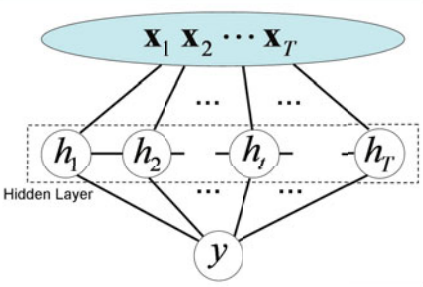


Fig. 2. Graphical representation of H-CRF. Our new model H-CORF (Sec. 3) shares the same structure. The upper two layers form CRF (and CORF in Sec. 2.3) when $\mathbf{h} = h_1, \ldots, h_T$ serves as observed outputs.

classical CRF states. We will refer to this model as the Conditional Ordinal Random Field (CORF). To develop the model we first introduce the framework of static ordinal regression and subsequently show how it can be extended into a structured, sequence setting.

2.1 Static Ordinal Regression

The goal of ordinal regression is to predict the label h of an item represented by a feature vector[1] $\mathbf{x} \in \mathbb{R}^p$ where the output indicates the preference or order of this item. Formally, we let $h \in \{1, \ldots, R\}$ for which R is the number of preference grades, and h takes an ordinal scale from the lowest preference $h = 1$ to the highest $h = R$, $h = 1 \prec h = 2 \prec \ldots \prec h = R$.

The most critical aspect that differentiates the ordinal regression approaches from the multi-class classification methods is the modeling strategy. Assuming a linear model (straightforwardly extendible to a nonlinear version by kernel tricks), the multi-class classification typically (c.f. [3]) takes the form of[2]

$$h = \arg\max_{c \in \{1,\ldots,R\}} \mathbf{w}_c^\top \mathbf{x} + b_c. \tag{1}$$

For each class c, the hyperplane ($\mathbf{w}_c \in \mathbb{R}^p, b_c \in \mathbb{R}$) defines the confidence toward the class c. The class decision is made by selecting the one with the largest confidence. The model parameters are $\{\{\mathbf{w}_c\}_{c=1}^R, \{b_c\}_{c=1}^R\}$. On the other hand, ordinal regression approaches adopt the following modeling strategy:

$$h = c \;\text{ iff }\; \mathbf{w}^\top \mathbf{x} \in (b_{c-1}, b_c], \quad \text{where } -\infty = b_0 \leq b_1 \leq \cdots \leq b_R = +\infty. \tag{2}$$

The binning parameters $\{b_c\}_{c=0}^R$ form R different bins, where their adjacent placement and the output deciding protocol of (2) naturally enforce the ordinal scale criteria. The parameters of the model become $\{\mathbf{w}, \{b_c\}_{c=0}^R\}$, far fewer

[1] We use the notation $\mathbf{x}$ interchangeably for both a sequence observation $\mathbf{x} = \{\mathbf{x}_r\}$ and a vector, which is clearly distinguished by context.

[2] This can be seen as a general form of the popular one-vs-all or one-vs-one treatment for the multi-class problem.

in count than those of the classification models. The state-of-the-art Support Vector Ordinal Regression (SVOR) algorithms [14,2] conform to this representation while they aim to maximize margins at the nearby bins in the SVM-like formulation.

2.2 Conditional Random Field (CRF) for Sequence Segmentation

CRF [10,9] is a structured output model which represents the distribution of a set (sequence) of categorical tags $\mathbf{h} = \{h_r\}, h_r \in \{1, \ldots, R\}$, conditioned on input $\mathbf{x}$. More formally, the density $P(\mathbf{h}|\mathbf{x})$ has a Gibbs form clamped on the observation $\mathbf{x}$:

$$P(\mathbf{h}|\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}; \boldsymbol{\theta})} e^{s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta})}. \tag{3}$$

Here $Z(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{h} \in \mathcal{H}} e^{s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta})}$ is the partition function on the space of possible configurations $\mathcal{H}$, and $\boldsymbol{\theta}$ are the parameters[3] of the *score function* $s(\cdot)$.

The choice of the output graph $G = (V, E)$ on $\mathbf{h}$ critically affects model's representational capacity and the inference complexity. For convenience, we further assume that we have either *node* cliques ($r \in V$) or *edge* cliques ($e = (r, s) \in E$) with corresponding features, $\boldsymbol{\Psi}_r^{(V)}(\mathbf{x}, h_r)$ and $\boldsymbol{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s)$. By letting $\boldsymbol{\theta} = \{\mathbf{v}, \mathbf{u}\}$ be the parameters for node and edge features, respectively, the score function is typically defined as:

$$s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta}) = \sum_{r \in V} \mathbf{v}^\top \boldsymbol{\Psi}_r^{(V)}(\mathbf{x}, h_r) + \sum_{e=(r,s) \in E} \mathbf{u}^\top \boldsymbol{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s). \tag{4}$$

In conventional modeling practice, the node/edge features are often defined as products of measurement features confined to cliques and the output class indicators. For instance, in CRFs with sequence [10] and lattice outputs [9,17] we often have

$$\boldsymbol{\Psi}_r^{(V)}(\mathbf{x}, h_r) = \Big[I(h_r = 1), \cdots, I(h_r = R) \Big]^\top \otimes \boldsymbol{\phi}(\mathbf{x}_r), \tag{5}$$

where $I(\cdot)$ is the indicator function and $\otimes$ denotes the Kronecker product. Hence the k-th block ($k = 1, \ldots, R$) of $\boldsymbol{\Psi}_r^{(V)}(\mathbf{x}, h_r)$ is $\boldsymbol{\phi}(\mathbf{x}_r)$ if $h_r = k$, and the $\mathbf{0}$-vector otherwise. The edge feature may typically assess the absolute difference between the measurements at adjoining nodes,

$$\Big[I(h_r = k \ \wedge \ h_s = l) \Big]_{R \times R} \otimes \big| \boldsymbol{\phi}(\mathbf{x}_r) - \boldsymbol{\phi}(\mathbf{x}_s) \big|. \tag{6}$$

Learning and inference in CRFs has been studied extensively in the past decade, c.f. [10,9,17], with many efficient and scalable algorithms, particularly for sequential structures.

[3] For brevity, we often drop the dependency on $\boldsymbol{\theta}$ in our notation.

2.3 Conditional Ordinal Random Field (CORF)

A standard CRF model seeks to *classify*, treating each output category nominally and equally different from all other categories. The consequence is that the model's node potential has a direct analogy to the static multi-class classification model of (1): For $h_r = c$, the node potential equals $\mathbf{v}_c^\top \boldsymbol{\phi}(\mathbf{x}_r)$ where $\mathbf{v}_c$ is the c-th block of $\mathbf{v}$, or the c-th hyperplane $\mathbf{w}_c^\top \mathbf{x}_r + b_c$ in (1). The max can be replaced by the *softmax* function. To setup an exact equality, one can let $\boldsymbol{\phi}(\mathbf{x}_r) = [1, \mathbf{x}_r^\top]^\top$.

Conversely, the modeling strategy of the static ordinal regression methods such as (2) can be merged with the CRF through the node potentials to yield a structured output ranking model. However, the mechanism of doing so is not obvious because of the highly discontinuous nature of (2). Instead, we base our approach on the probabilistic model for ranking proposed by [1], which shares the notion of (2).

In [1], the noiseless probabilistic ranking likelihood is defined as

$$P_{ideal}(h = c|f(\mathbf{x})) = \begin{cases} 1 \text{ if } f(\mathbf{x}) \in (b_{c-1}, b_c] \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

Here $f(\mathbf{x})$ is the model to be learned, which could be linear $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. The effective ranking likelihood is constructed by contaminating the ideal model with noise. Under the Gaussian noise δ and after marginalization, one arrives at the ranking likelihood

$$P(h=c|f(\mathbf{x})) = \int_\delta P_{ideal}(h=c|f(\mathbf{x})+\delta)\cdot\mathcal{N}(\delta;0,\sigma^2)d\delta = \Phi\left(\frac{b_c - f}{\sigma}\right) - \Phi\left(\frac{b_{c-1} - f}{\sigma}\right), \tag{8}$$

where $\Phi(\cdot)$ is the standard normal cdf, and σ is the parameter that controls the steepness of the likelihood function.

Now we set the node potential at node r of the CRF to be the log-likelihood of (8), that is,

$$\begin{aligned} &\mathbf{v}^\top \boldsymbol{\Psi}_r^{(V)}(\mathbf{x}, h_r) \longrightarrow \boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r; \{\mathbf{a}, \mathbf{b}, \sigma\}), \quad \text{where} \\ &\boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r) := \textstyle\sum_{c=1}^R I(h_r = c) \cdot \log\left(\Phi\left(\frac{b_c - \mathbf{a}^\top \boldsymbol{\phi}(\mathbf{x}_r)}{\sigma}\right) - \Phi\left(\frac{b_{c-1} - \mathbf{a}^\top \boldsymbol{\phi}(\mathbf{x}_r)}{\sigma}\right)\right). \end{aligned} \tag{9}$$

Here, $\mathbf{a}$ (having the same dimension as $\boldsymbol{\phi}(\mathbf{x}_r)$), $\mathbf{b} = [-\infty = b_0, \ldots, b_R = +\infty]^\top$, and σ are the new parameters, in contrast with the original CRF's node parameters $\mathbf{v}$. Substituting this expression into (4) leads to a new conditional model for structured ranking,

$$P(\mathbf{h}|\mathbf{x}, \boldsymbol{\omega}) \propto \exp\big(s(\mathbf{x}, \mathbf{h}; \boldsymbol{\omega})\big), \text{ where} \tag{10}$$

$$s(\mathbf{x}, \mathbf{h}; \boldsymbol{\omega}) = \sum_{r \in V} \boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r; \{\mathbf{a}, \mathbf{b}, \sigma\}) + \sum_{e=(r,s)\in E} \mathbf{u}^\top \boldsymbol{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s). \tag{11}$$

We refer to this model as *CORF*, the Conditional Ordinal Random Field. The parameters of the CORF are denoted as $\boldsymbol{\omega} = \{\mathbf{a}, \mathbf{b}, \sigma, \mathbf{u}\}$, with the ordering

constraint $b_i < b_{i+1}, \forall i$. Note that the number of parameters is significantly fewer than that of the regular CRF. Unlike CRF's log-linear form, the CORF becomes a *log-nonlinear* model, effectively imposing the ranking criteria via nonlinear binning-based modeling of the node potential $\boldsymbol{\Gamma}$.

Model Learning. We briefly discuss how the CORF model can be learned using gradient ascent. For the time being we assume that we are given labeled data pairs $(\mathbf{x}, \mathbf{h})$, a typical setting for CRF learning, although we treat $\mathbf{h}$ as latent variables for the H-CORF sequence classification model in Sec. 3.

First, it should be noted that CORF's log-nonlinear modeling does not impose any additional complexity on the inference task. Since the graph topology remains the same, once the potentials are evaluated, the inference follows exactly the same procedures as that of the standard log-linear CRFs. Second, it is not difficult to see that the node potential $\boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r)$, although non-linear, remains concave.

Unfortunately, the overall learning of CORF is non-convex because of the log-partition function (*log-sum-exp* of nonlinear concave functions). However, the log-likelihood objective is bounded above by 0, and the quasi-Newton or the stochastic gradient ascent [17] can be used to estimate the model parameters. The gradient of the log-likelihood w.r.t. $\mathbf{u}$ is (the same as the regular CRF):

$$\frac{\partial \log P(\mathbf{h}|\mathbf{x}, \boldsymbol{\omega})}{\partial \mathbf{u}} = \sum_{e=(r,s)\in E} \left(\boldsymbol{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s) - \mathbb{E}_{P(h_r,h_s|\mathbf{x})} \left[\boldsymbol{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s) \right] \right). \tag{12}$$

The gradient of the log-likelihood w.r.t. $\mu = \{\mathbf{a}, \mathbf{b}, \sigma\}$ can be derived as:

$$\frac{\partial \log P(\mathbf{h}|\mathbf{x}, \boldsymbol{\omega})}{\partial \mu} = \sum_{r\in V} \left(\frac{\partial \boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r)}{\partial \mu} - \mathbb{E}_{P(h_r|\mathbf{x})} \left[\frac{\partial \boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r)}{\partial \mu} \right] \right), \tag{13}$$

where the gradient of the node potential can be computed analytically,

$$\frac{\partial \boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r)}{\partial \mu} = \sum_{c=1}^{R} I(h_r{=}c) \cdot \frac{\mathcal{N}(z_0(r,c); 0, 1) \cdot \frac{\partial z_0(r,c)}{\partial \mu} - \mathcal{N}(z_1(r,c); 0, 1) \cdot \frac{\partial z_1(r,c)}{\partial \mu}}{\Phi(z_0(r,c)) - \Phi(z_1(r,c))},$$

$$\text{where } z_k(r,c) = \frac{b_{c-k} - \mathbf{a}^\top \boldsymbol{\phi}(\mathbf{x}_r)}{\sigma} \text{ for } k = 0, 1. \tag{14}$$

Model Reparameterization for Unconstrained Optimization. The gradient-based learning proposed above has to be accomplished while respecting two sets of constraints: (i) the order constraints on $\mathbf{b}$: $\{b_{j-1} \leq b_j$ for $j = 1, \ldots, R\}$, and (ii) the positive scale constraint on σ: $\{\sigma > 0\}$. Instead of general constrained optimization, we introduce a reparameterization that effectively reduces the problem to an unconstrained optimization task.

To deal with the order constraints in the parameters $\mathbf{b}$, we introduce the displacement variables δ_k, where $b_j = b_1 + \sum_{k=1}^{j-1} \delta_k^2$ for $j = 2, \ldots, R-1$. So, $\mathbf{b}$

is replaced by the unconstrained parameters $\{b_1, \delta_1, \ldots, \delta_{R-2}\}$. The positiveness constraint for σ is simply handled by introducing the free parameter σ_0 where $\sigma = \sigma_0^2$. Hence, the unconstrained node parameters are: $\{\mathbf{a}, b_1, \delta_1, \ldots, \delta_{R-2}, \sigma_0\}$. Then the gradients for $\frac{\partial z_k(r,c)}{\partial \mu}$ in (14) then become:

$$\frac{\partial z_k(r,c)}{\partial \mathbf{a}} = -\frac{1}{\sigma_0^2}\boldsymbol{\phi}(\mathbf{x}_r), \quad \frac{\partial z_k(r,c)}{\partial \sigma_0} = -\frac{2\big(b_{c-k} - \mathbf{a}^\top \boldsymbol{\phi}(\mathbf{x}_r)\big)}{\sigma_0^3}, \quad \text{for } k = 0, 1. \tag{15}$$

$$\frac{\partial z_0(r,c)}{\partial b_1} = \begin{cases} 0 & \text{if } c = R \\ \frac{1}{\sigma_0^2} & \text{otherwise} \end{cases}, \quad \frac{\partial z_1(r,c)}{\partial b_1} = \begin{cases} 0 & \text{if } c = 1 \\ \frac{1}{\sigma_0^2} & \text{otherwise} \end{cases}. \tag{16}$$

$$\frac{\partial z_0(r,c)}{\partial \delta_j} = \begin{cases} 0 & \text{if } c \in \{1, \ldots, j, R\} \\ \frac{2\delta_j}{\sigma_0^2} & \text{otherwise} \end{cases}, \quad \frac{\partial z_1(r,c)}{\partial \delta_j} = \begin{cases} 0 & \text{if } c \in \{1, \ldots, j+1\} \\ \frac{2\delta_j}{\sigma_0^2} & \text{otherwise} \end{cases}, \quad \text{for } j = 1, \ldots, R-2. \tag{17}$$

We additionally employ parameter regularization on the CORF model. For $\mathbf{a}$ and $\mathbf{u}$, we use the typical L2 regularizers $||\mathbf{a}||^2$ and $||\mathbf{u}||^2$. No specific regularization is necessary for the binning parameters b_1 and $\{\delta_j\}_{j=1}^{R-2}$ as they will be automatically adjusted according to the score $\mathbf{a}^\top \boldsymbol{\phi}(\mathbf{x}_r)$. For the scale parameter σ_0 we consider $(\log \sigma_0^2)^2$ as the regularizer, which essentially favors $\sigma_0 \approx 1$ and imposes quadratic penalty in log-scale.

3 Hidden Conditional Ordinal Random Field (H-CORF)

We now propose an extension of the CORF model to a sequence classification setting. The model builds upon the method for extending CRFs for classification, known as Hidden CRFs (H-CRF). H-CRF is a probabilistic classification model $P(y|\mathbf{x})$ that can be seen as a combination of K CRFs, one for each class. The CRF's output variables $\mathbf{h} = h_1, \ldots, h_T$ are now treated as latent variables (Fig. 2). H-CRF has been studied in the fields of computer vision [12,18] and speech recognition [5]. We use the same approach to combine individual CORF models as building blocks for sequence classification in the Hidden CORF setting, a structured ordinal regression model with latent variables.

To build a classification model from CORFs, we introduce a class variable $y \in \{1, \ldots, K\}$ and a new score function

$$\begin{aligned} s(y, \mathbf{x}, \mathbf{h}; \boldsymbol{\Omega}) &= \sum_{k=1}^{K} I(y = k) \cdot s(\mathbf{x}, \mathbf{h}; \boldsymbol{\omega}_k) \\ &= \sum_{k=1}^{K} I(y = k) \cdot \left[\sum_{r \in V} \boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r; \{\mathbf{a}_k, \mathbf{b}_k, \sigma_k\}) + \sum_{e=(r,s) \in E} \mathbf{u}_k^\top \boldsymbol{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s) \right], \end{aligned} \tag{18}$$

where $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_k\}_{k=1}^K$ denotes the compound H-CORF parameters comprised of K CORFs $\boldsymbol{\omega}_k = \{\mathbf{a}_k, \mathbf{b}_k, \sigma_k, \mathbf{u}_k\}$ for $k = 1, \ldots, K$. The score function, in turn, defines the joint and class conditional distributions:

$$P(y, \mathbf{h}|\mathbf{x}) = \frac{\exp(s(y, \mathbf{x}, \mathbf{h}))}{Z(\mathbf{x})}, \quad P(y|\mathbf{x}) = \sum_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{x}) = \frac{\sum_{\mathbf{h}} \exp(s(y, \mathbf{x}, \mathbf{h}))}{Z(\mathbf{x})}. \tag{19}$$

Evaluation of the class-conditional $P(y|\mathbf{x})$ depends on the partition function $Z(\mathbf{x}) = \sum_{y,\mathbf{h}} \exp(s(y, \mathbf{x}, \mathbf{h}))$ and the class-latent joint posteriors $P(y, h_r, h_s|\mathbf{x})$. Both can be computed from independent consideration of K individual CORFs. The compound partition function is the sum of individual partition functions, $Z(\mathbf{x}) = \sum_k Z(\mathbf{x}|y = k) = \sum_k \sum_{\mathbf{h}} \exp(s(k, \mathbf{x}, \mathbf{h}))$, computed in each CORF. Similarly, the joint posteriors can evaluated as $P(y, h_r, h_s|\mathbf{x}) = P(h_r, h_s|\mathbf{x}, y) \cdot P(y|\mathbf{x})$. Learning the H-CORF can be done by maximizing the class conditional log-likelihood $\log P(y|\mathbf{x})$, where its gradient can be derived as:

$$\frac{\partial \log P(y|\mathbf{x})}{\partial \boldsymbol{\Omega}} = \mathbb{E}_{P(\mathbf{h}|\mathbf{x},y)}\left[\frac{\partial s(y, \mathbf{x}, \mathbf{h})}{\partial \boldsymbol{\Omega}}\right] - \mathbb{E}_{P(y,\mathbf{h}|\mathbf{x})}\left[\frac{\partial s(y, \mathbf{x}, \mathbf{h})}{\partial \boldsymbol{\Omega}}\right]. \tag{20}$$

Using the gradient derivation (12)-(14) for the CORF, it is straightforward to compute the expectations in (20). Finally, the assignment of a measurement sequence to a particular class, such as the action or emotion, is accomplished by the MAP rule $y^* = \arg\max_y P(y|\mathbf{x})$.

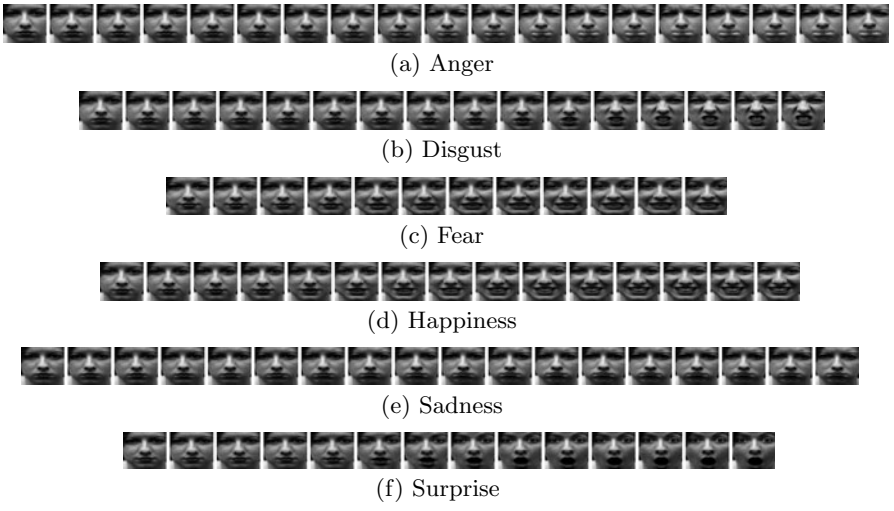
4 Evaluations

In this section we demonstrate the performance of our model with ordinal latent state dynamics, the H-CORF. We evaluate algorithms on two datasets/tasks: facial emotion recognition from the CMU facial expression video dataset and behavior recognition from the UCSD mouse dataset.

4.1 Recognizing Facial Emotions from Videos

We consider the task of the facial emotion recognition. We use the Cohn-Kanade facial expression database [11], which consists of six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) performed by 100 students, 18 to 30 years old. In this experiment, we selected image sequences from 93 subjects, each of which enacts 2 to 6 emotions. Overall, the number of sequences is 352 where the class proportions are as follows: anger(36), disgust(42), fear(54), happiness(85), sadness(61), and surprise(74). For this 6-class problem, we randomly select 60%/40% of the sequences as training/testing, respectively. The training and the testing sets do not have sequences of the same subject. After detecting faces with the cascaded face detector [16], we normalize them into (64×64) images which are aligned based on the eye locations similar to [15].

Unlike the previous static emotion recognition approaches (e.g., [13]) where just the ending few peak frames are considered, we use the entire sequences that cover the onset state of the expression to the apex in order to conduct the task of dynamic emotion recognition. The sequence lengths are, on average, about 20 frames long. Fig. 3 shows some example sequences. We consider the qualitative



(a) Anger

(b) Disgust

(c) Fear

(d) Happiness

(e) Sadness

(f) Surprise

Fig. 3. Sample sequences for six emotions from the Cohn-Kanade dataset

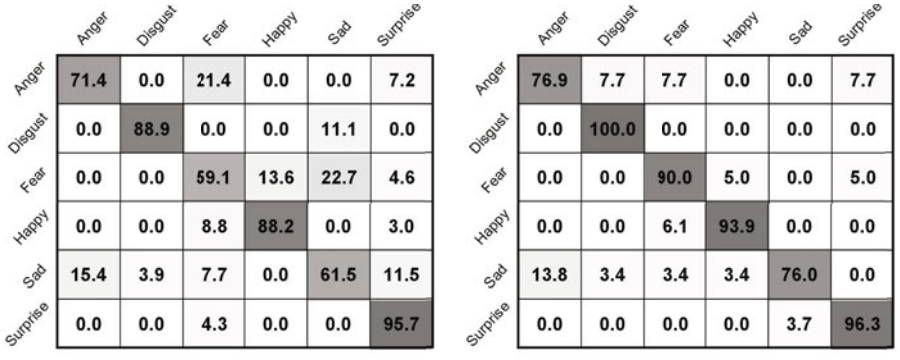
intensity state of size $R = 3$, based on typical representation of three ordinal categories used to describe the emotion dynamics: `neutral` $<$ `increasing` $<$ `apex`. Note that we impose no actual prior knowledge of the category dynamics nor the correspondence of the three states to the qualitative categories described above. This correspondence can be established by interpreting the model learned in the estimation stage, as we demonstrate next. For the image features, we first extract the Haar-like features, following [20]. To reduce feature dimensionality, we apply PCA on the training frames for each emotion, which gives rise to 30-dimensional feature vectors corresponding to 90% of the total energy.

The recognition test errors are shown in Table 1. Here we also contrasted with the baseline generative approach based on a Gaussian Hidden Markov Model (GHMM). See also the confusion matrices of H-CRF and H-CORF in Fig. 4. Our model with ordinal dynamics leads to significant improvements in classification performance over both prior models.

To gain insight about the modeling ability of the new approach, we studied the latent intensity envelopes learned during the model estimation phase. Fig. 5 depicts a set of most likely latent envelopes estimated on a sample of test sequences. The decoded envelopes by our model correspond to typical visual changes in the emotion intensities, qualified by the three categories (neutral, increase, apex). On the other hand, the decoded states by the H-CRF model have weaker correlation with the three target intensity categories, typically exhibiting highly diverse scales and/or orders across the six emotions. The ability of the ordinal model to recover perceptually distinct dynamic categories from data may further explain the model's good classification performance.

Table 1. Recognition accuracy on CMU emotion video dataset

Methods	GHMM	H-CRF	H-CORF
Accuracy	72.99%	78.10%	89.05%



(a) H-CRF (b) (Proposed) H-CORF

Fig. 4. Confusion matrices for facial emotion recognition on CMU database

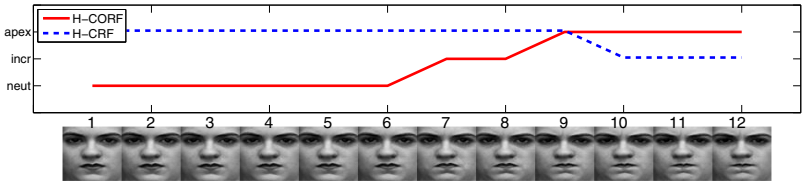
4.2 Behavior Recognition from UCSD Mouse Dataset

We next consider the task of behavior recognition from video, a very important problem in computer vision. We used the mouse dataset from the UCSD vision group[4]. The dataset contains videos of 5 different mouse behaviors (drink, eat, explore, groom, and sleep). See Fig. 6 for some sample frames. The video clips are taken at 7 different points in the day, separately kept as 7 different sets. The characteristics of each behavior vary substantially among each of the seven sets. From the original dataset, we select a subset comprised of 75 video clips (15 videos for each behavior) from 5 sets. Each video lasts between 1 and 10 seconds. For the recognition setting, we take one of the 5 sets having the largest number of instances (25 clips; 5 for each class) as the training set, while the remaining 50 videos from the other 4 sets are reserved for testing.

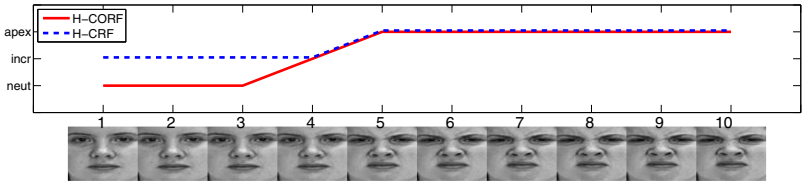
To obtain the measurement features from the raw videos, we extract dense spatio-temporal 3D *cuboid* features of [4]. Similar to [4], we construct a finite codebook of descriptors, and replace each cuboid descriptor by the corresponding codebook word. More specifically, after collecting the cuboid features from all videos, we cluster them into $C = 200$ centers using the k-means algorithm.

For the baseline performance comparison, we first run [4]'s static mixture approach where each video is represented as a static histogram of cuboid types contained in the video clip, essentially forming a bag-of-words representation. We then apply standard classification methods such as the nearest neighbor (NN)

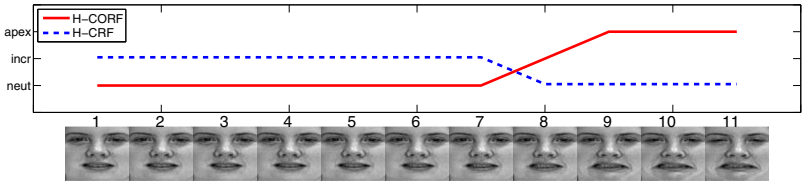
[4] Available for download at http://vision.ucsd.edu



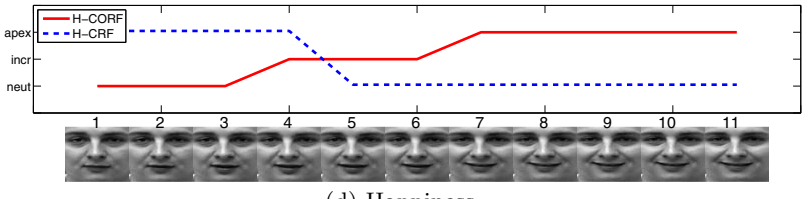
(a) Anger



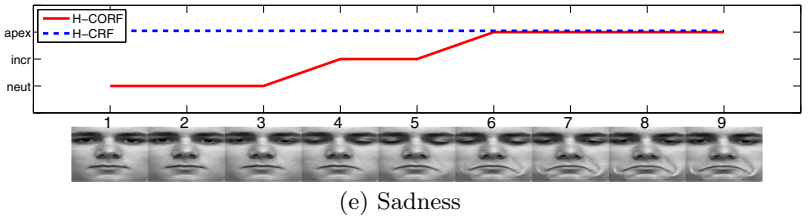
(b) Disgust



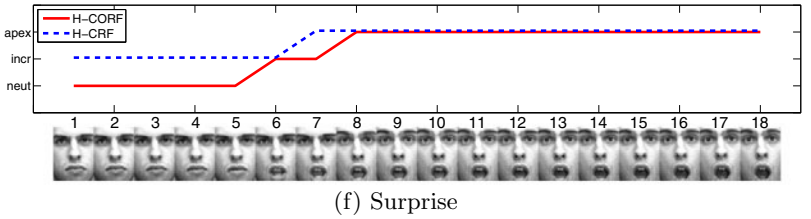
(c) Fear



(d) Happiness



(e) Sadness



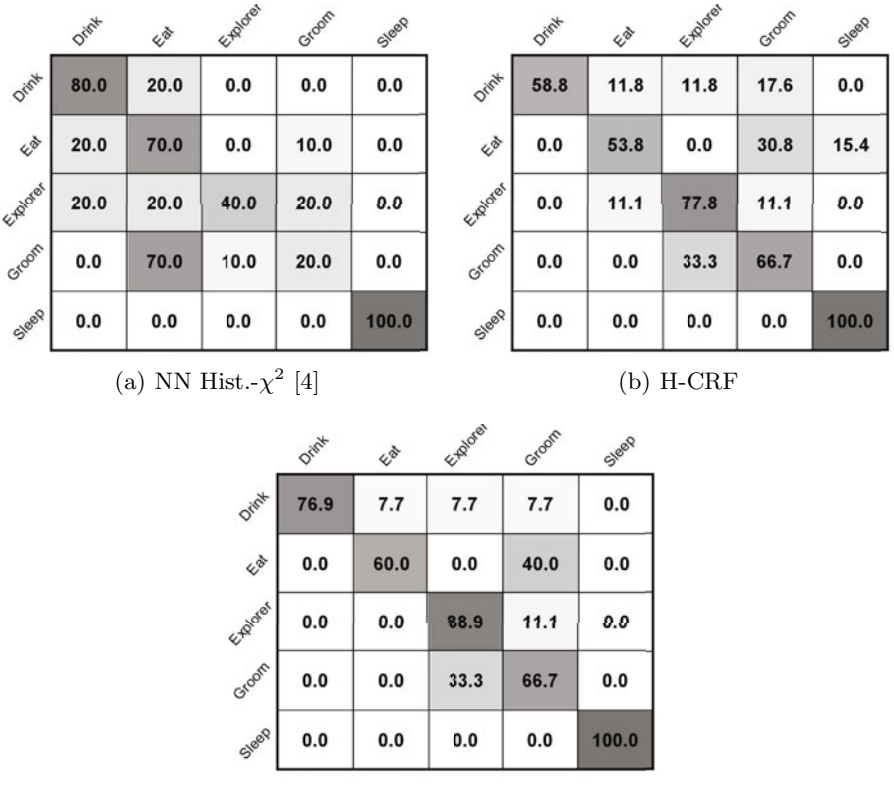
(f) Surprise

Fig. 5. Facial emotion intensity prediction for some test sequences. The decoded latent states by H-CORF are shown as red lines, contrasted with H-CRF's blue dotted lines.

Fig. 6. Sample frames from mouse dataset, representing each of the five classes (drink, eat, explore, groom, and sleep) from left to right

Table 2. Recognition accuracy on UCSD mouse dataset

Methods	NN Hist.-χ^2 [4]	GHMM	H-CRF	H-CORF
Accuracy	62.00%	64.00%	68.00%	78.00%



(a) NN Hist.-χ^2 [4]

(b) H-CRF

(c) (Proposed) H-CORF

Fig. 7. Confusion matrices for behavior recognition in UCSD mouse dataset

classifier based on the χ^2 distance measure on the histogram space. We obtain the test accuracy (Table 2) and the confusion matrix (Fig. 7) shown under the title "NN Hist.-χ^2". Note that the random guess would yield 20.00% accuracy.

Instead of representing the video as a single histogram, we consider a sequence representation for our H-CORF-based sequence models. For each time frame t, we set a time-window of size $W = 40$ centered at t. We then collect all detected cuboids with the window, and form a histogram of cuboid types as the node feature $\phi(\mathbf{x}_r)$. Note that some time slices may have no cuboids involved, in which case the feature vector is a zero-vector. To avoid a large number of parameters in the learning, we further reduce the dimensionality of features to 100-dim by PCA which corresponds to about 90% of the total energy.

The test errors and the confusion matrices of the H-CRF and our H-CORF are contrasted with the baseline approach in Table 2 and Fig. 7. Here the cardinality of the latent variables is set as $R = 3$ to account for different ordinal intensity levels of mouse motions, which is chosen among a set of values that produced highest prediction accuracy. Our H-CORF exhibits better performance than the H-CRF and [4]'s standard histogram-based approach. Results similar to ours have been reported in other works that use more complex models and are evaluated on the same dataset (c.f., [19]). However, they are not immediately comparable to ours as we have different experimental settings: a smaller subset with non-overlapping sessions (i.e., sets) between training and testing where we have a much smaller training data proportion (33.33%) than [19]'s (63.33%).

5 Conclusion

In this paper we have introduced a new modeling framework of Hidden Conditional Ordinal Random Fields to accomplish the task of sequence classification. The H-CORF, by introducing a set of ordinal-scale latent variables, aims at modeling the qualitative intensity envelope constraints often observed in real human/animal motions. The embedded sequence segmentation model, CORF, extends the regular CRF by incorporating the ranking-based potentials to model dynamically changing ordinal-scale signals. For the real datasets for facial emotion and mouse behavior recognition, we have demonstrated that the faithful representation of the linked ordinal states in our H-CORF is highly useful for accurate classification of entire sequences. In our future work, we will apply our method to more extensive and diverse types of sequence datasets including biological and financial data.

Acknowledgments. We are grateful to Peng Yang and Dimitris N. Metaxas for their help and discussions throughout the course of this work. This material is based upon work supported by the National Science Foundation under Grant No. IIS-0916812.

References

[1] Chu, W., Ghahramani, Z.: Gaussian processes for ordinal regression. Journal of Machine Learning Research 6, 1019–1041 (2005)

[2] Chu, W., Keerthi, S.S.: New approaches to support vector ordinal regression. In: International Conference on Machine Learning (2005)

[3] Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research 2, 265–292 (2001)
[4] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (2005)
[5] Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C.: Hidden conditional random fields for phone classification. In: International Conference on Speech Communication and Technology (2005)
[6] Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. In: Advances in Large Margin Classifiers. MIT Press, Cambridge (2000)
[7] Hu, Y., Li, M., Yu, N.: Multiple-instance ranking: Learning to rank images for image retrieval. In: Computer Vision and Pattern Recognition (2008)
[8] Jing, Y., Baluja, S.: Pagerank for product image search. In: Proceeding of the 17th international conference on World Wide Web (2008)
[9] Kumar, S., Hebert, M.: Discriminative random fields. International Journal of Computer Vision 68, 179–201 (2006)
[10] Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (2001)
[11] Lien, J., Kanade, T., Cohn, J., Li, C.: Detection, tracking, and classification of action units in facial expression. Journal of Robotics and Autonomous Systems (1999)
[12] Quattoni, A., Collins, M., Darrell, T.: Conditional random fields for object recognition. In: Neural Information Processing Systems (2004)
[13] Shan, C., Gong, S., McOwan, P.W.: Conditional mutual information based boosting for facial expression recognition. In: British Machine Vision Conference (2005)
[14] Shashua, A., Levin, A.: Ranking with large margin principle: Two approaches. In: Neural Information Processing Systems (2003)
[15] Tian, Y.: Evaluation of face resolution for expression analysis. In: Computer Vision and Pattern Recognition Workshop on Face Processing in Video (2004)
[16] Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision 57(2), 137–154 (2001)
[17] Vishwanathan, S., Schraudolph, N., Schmidt, M., Murphy, K.: Accelerated training of conditional random fields with stochastic meta-descent. In: International Conference on Machine Learning (2006)
[18] Wang, S., Quattoni, A., Morency, L.P., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: Computer Vision and Pattern Recognition (2006)
[19] Willems, G., Tuytelaars, T., Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
[20] Yang, P., Liu, Q., Metaxas, D.N.: Rankboost with l1 regularization for facial expression recognition and intensity estimation. In: International Conference on Computer Vision (2009)