Multi-output Laplacian Dynamic Ordinal Regression for Facial Expression Recognition and Intensity Estimation

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

Abstract

Automated facial expression recognition has received increased attention over the past two decades. Existing works in the field usually do not encode either the temporal evolution or the intensity of the observed facial displays. They also fail to jointly model multidimensional (multi-class) continuous facial behaviour data; binary classifiers - one for each target basic-emotion class - are used instead. In this paper, intrinsic topology of multidimensional continuous facial affect data is first modeled by an ordinal manifold. This topology is then incorporated into the Hidden Conditional Ordinal Random Field (H-CORF) framework for dynamic ordinal regression by constraining H-CORF parameters to lie on the ordinal manifold. The resulting model attains simultaneous dynamic recognition and intensity estimation of facial expressions of multiple emotions. To the best of our knowledge, the proposed method is the first one to achieve this on both deliberate as well as spontaneous facial affect data.

1. Introduction

Facial behavior is believed to be the most important source of information when it comes to affect, attitude, intentions, and social signals interpretation [2]. Automatic facial expression recognition has therefore been an active topic of research for more than two decades [17, 25]. Most systems developed so far attempt automatic recognition of prototypic facial expressions of six basic emotions (anger, happiness, fear, surprise, sadness, and disgust). The main criticism that these works received from both cognitive and computer scientists is that the methods are not applicable in real-world situations, where subtle changes in both appearance and temporal evolution of facial expressions typify the displayed facial behavior [2, 1]. Current works in the field usually do not encode either the intensity of the observed facial appearance changes or the evolution of these changes in time [25]. Instead, current approaches usually apply six binary classifiers - one for each target prototypic facial expression of emotion - that code input face imagery as either belonging to the target class or not.

Exceptions to this trend include a small number of works on automatic coding of facial imagery in terms of either temporal segments of facial actions (e.g., [22, 11, 14, 18]), or temporal segments of prototypic expressions of emotions (e.g., [7, 10]), or a small number of prototypic facial expression intensity levels (e.g., [5]). Some of the past works in the field have proposed methods that could be used for recognition of facial expression temporal segments and/or intensity levels (e.g., [23, 20]), but did not actually report any quantitative results for that task. Most of these works use temporal graphical models being either generative (e.g., Hidden Markov Models (HMM), [7, 22, 11]) or discriminative (e.g., CRFs [10]) trained for recognition of temporal segments of a target facial expression. However, most of these approaches fail to jointly model different emotions, making the models suboptimal for the emotion modeling task.

A method that does not conform to this rule is the H-CORF model [9], which has been successfully used for simultaneous recognition of multiple emotion-related expressions and their intensities. Yet, despite improvements over other dynamic models (e.g., HMM or standard CRF), H-CORF relies on linear feature models. Such 'simple' feature representation is usually not discriminative enough for recognition and intensity estimation of facial behaviour due to the large variation in expressions and their intensity among different subjects.

In this paper, we propose to model topology of the input data by a low-dimensional manifold that preserves discriminative information about various facial expressions of

emotions and ordinal relationships between their intensities while being largely invariant to intra- and inter-subject variations. We incorporate this topology into the H-CORF framework for dynamic ordinal regression by constraining H-CORF parameters to lie on this nonlinear manifold. To keep the model computationally tractable we adopt a locally linear approximation of the otherwise nonlinear parameter manifold. This manifold approximation is then coupled to and *jointly estimated* with the H-CORF. In this manner we directly find the most discriminative features for dynamic recognition of emotions and their intensities. To the best of our knowledge, this is the first method that performs simultaneous recognition of multiple facial expressions of emotions and their intensities by modeling all: (i) temporal dynamics of facial expressions, (ii) ordinal relationships between their intensities, and (iii) intrinsic topology of multidimensional continuous facial affect data, encoded by an ordinal manifold structure.

In what follows, we consider a K-class classification problem, where we let $c \in \{1, ..., K\}$ be the nominal category (i.e., emotion class). Each nominal category c is assumed to have R different ordinal scales (i.e., emotions intensities), denoted as consecutive integers $h_r \in \{1, \dots, R\}$ that keep the ordering information. The observations, denoted by $\mathbf{x} = \mathbf{x}_1 \dots \mathbf{x}_T$ and where the sequence length T can vary from instance to instance, serve as input covariates for predicting both c and h. If not stated otherwise, we assume a fully supervised setting: we are given a training set of N data triplets $\mathcal{D} = \{(c^i, h^i, \mathbf{x}^i)\}_{i=1}^N$, which are i.i.d. samples from an underlying but unknown distribution.

The remainder of the paper is organized as follows. We give a short overview of of the models for dynamic ordinal regression in Sec. 2. We describe the manifold learning approach employed in the proposed model in Sec. 3. The proposed model for dynamic ordinal regression is described in Sec. 4. Sec. 5 shows the experimental results. Sec. 6 concludes the paper.

2. Dynamic 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: $h = 1 \prec h = 2 \prec \dots \prec h = R$. Modeling of the item orders can be accomplished by means of standard static ordinal regression models (e.g [21, 4, 3]), which, in contrast to multi-class classification models, preserve ordering relationships between different labels. Nevertheless, static models for ordinal regression ignore temporal correlations between the labels, which is of essence when dealing with sequential data. In what follows, we describe two recently proposed dynamic models for ordinal regression [10, 9].

[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.

2.1. Conditional Ordinal Random Field (CORF)

CORF [10] is an extension of standard CRF [12] to structured output ordinal regression setting. It models the distribution of a set (sequence) of ordinal variables $\mathbf{h} = \{h_r\}, h_r \in \{1, \dots, R\}$, conditioned on inputs $\mathbf{x}$. As in standard CRF, the distribution $P(\mathbf{h}|\mathbf{x})$ has a Gibbs form clamped on the observation $\mathbf{x}$:

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

where $Z(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{h} \in \mathcal{H}} e^{s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta})}$ is the normalizing partition function ($\mathcal{H}$ is a set of all possible output configurations), and $\boldsymbol{\theta}$ are the parameters[2] of the score function $s(\cdot)$, defined as

$$\begin{aligned} s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta}) = &\sum_{r \in V} \boldsymbol{\Gamma}_r^{(V)}(\mathbf{x}, h_r; \{\mathbf{a}, \mathbf{b}, \sigma\}) \\ &+ \sum_{e=(k,l) \in E} \mathbf{u}_{k,l}^\top \boldsymbol{\Psi}_e^{(E)}(\mathbf{x}, h_r = k, h_s = l), \end{aligned} \tag{2}$$

thus, summing up the influence of the node features (Ψ_r) and the edge features (Ψ_e) on the model output. However, in contrast to standard CRF, CORF employs the modeling strategy of static ordinal regression methods (see [3]) to define the node features Ψ_r. Specifically, the probabilistic ranking likelihood, $P(h = c|f(x)) = P(f(x) \in [b_{c-1}, b_c))$, is used, where $f(\mathbf{x}) = \mathbf{a}^\top \boldsymbol{\phi}(\mathbf{x})$ is the *linear* model in the induced feature space $\boldsymbol{\phi}(\mathbf{x})$. Thus, a projects the features $\boldsymbol{\phi}(\mathbf{x})$ on a single line divided into R bins, with the binning parameters $\mathbf{b} = [-\infty = b_0, \dots, b_R = +\infty]^\top$, which satisfy the ordinal constraints ($b_i < b_{i+1}, \forall i$). Under the Gaussian noise assumption, the ranking likelihood becomes

$$P(h = c|f(\mathbf{x})) = \Phi\left(\frac{b_c - f(\mathbf{x})}{\sigma}\right) - \Phi\left(\frac{b_{c-1} - f(\mathbf{x})}{\sigma}\right), \tag{3}$$

where $\Phi(\cdot)$ is the standard normal cdf, and σ is the parameter that controls the steepness of the likelihood function [3]. The ranking likelihood in (3) is used to set the node potential at node r in the CORF model as $\boldsymbol{\Psi}_r^{(V)}(\mathbf{x}, h_r = c) = \log P(h = c|f(\mathbf{x}))$, while the edge features, $\boldsymbol{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s)$, are set as

$$\Big[I(h_r = k \ \wedge \ h_s = l)\Big]_{R \times R} \otimes \big|\boldsymbol{\phi}(\mathbf{x}_r) - \boldsymbol{\phi}(\mathbf{x}_s)\big|. \tag{4}$$

$I(\cdot)$ is the indicator function that returns 1 (0) if the argument is true (false), and $\otimes$ denotes the Kronecker product. Finally, the parameters of the CORF model are stored in $\boldsymbol{\theta} = \{\mathbf{a}, \mathbf{b}, \sigma, \mathbf{u}\}$, and $\boldsymbol{\phi}(\mathbf{x}) = [1, x^T]^T$, as in [10] .

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

2.2. Multi-output CORF (M-CORF)

The output of the CORF model introduced in the previous section comprises ordinal variables $h_r \in \{1 \dots R\}$ corresponding to a single class. To deal with multiple classes, $c = \{1, ..., K\}$, H-CORF [9] combines K independent CORF models by employing the modeling strategy of Hidden CRF (H-CRF) [16], resulting in a new score function

$$s(c, \mathbf{x}, \mathbf{h}; \mathbf{\Omega}) = \sum_{k=1}^{K} I(c = k) \cdot s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta}_k) \tag{5}$$

where $s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta}_k)$ is defined by (2), and $\mathbf{\Omega} = \{\boldsymbol{\theta}_k\}_{k=1}^{K}$, where $\boldsymbol{\theta}_k = \{\mathbf{a}_k, \mathbf{b}_k, \sigma_k, \mathbf{u}_k\}$ for $k = 1 \dots K$, are the parameters of H-CORF. With the new score function, the joint and class conditional distributions are given by

$$P(c, \mathbf{h}|\mathbf{x}) = \frac{\exp(s(c, \mathbf{x}, \mathbf{h}))}{Z(\mathbf{x})}. \tag{6}$$

$$P(c|\mathbf{x}) = \sum_{\mathbf{h}} P(c, \mathbf{h}|\mathbf{x}) = \frac{\sum_{\mathbf{h}} \exp(s(c, \mathbf{x}, \mathbf{h}))}{Z(\mathbf{x})} \tag{7}$$

Evaluation of the class-conditional $P(c|\mathbf{x})$ depends on the partition function $Z(\mathbf{x}) = \sum_{c,\mathbf{h}} \exp(s(c, \mathbf{x}, \mathbf{h}))$ and the class-latent joint posteriors $P(c, h_r, h_s|\mathbf{x})$. Both can be computed from independent consideration of K individual CORFs. Note that the H-CORF model treats ordinal variables h as latent variables, and, thus, does not employ the corresponding labels during training. In what follows, we also consider a fully supervised setting in which labels for both classes c and ordinal variables h are known. To distinguish this setting from standard H-CORF, we call it Multi-output CORF (M-CORF).

Shared-paramater M-CORF: SM-CORF

In M-CORF, each CORF component is assigned an independent set of parameters ($\boldsymbol{\theta}_k$). However, since our classes are related[3] , it seems natural to use some shared parameters that 'couple' individual CORF components so that similarities across them can be exploited. Furthermore, the parameter sharing should constrain the parameters to a more plausible region of the parameter space. This is achieved by modeling intensities of different emotions on a common real line, divided by the binning parameters $\mathbf{b}$, which are shared among all classes (emotions). We call this model the Shared-parameter M-CORF (SM-CORF) model, where a set of parameters $\{\mathbf{b}, \sigma\}$ is shared among all CORF components, while ordinal projections $\mathbf{a_k}$ and transition matrix $\mathbf{u_k}$ are emotion-specific. In the same way, we define the Shared parameter H-CORF (SH-CORF) model.

[3]This comes from the fact that temporal segments of each emotion class can be labeled as neutral, onset and apex, where, e.g. neutral should be the same for all emotions

3. Manifold for Ordinal Regression

The goal of standard manifold learning is to discover a latent space in which topology of the input features $\mathbf{x}$, sometimes also informed by labels of $\mathbf{x}$, is preserved. Such data representation may be more discriminative and better suited for modeling of dynamic ordinal regression. In what follows, we first describe an unsupervised method for manifold learning. We then extend this method to obtain a manifold that satisfies ordinal constraints. Finally, we show how this ordinal manifold can be incorporated into the HCORF framework for dynamic ordinal regression.

3.1. Locality Preserving Projection (LPP)

Locality Preserving Projection (LPP) [8] is the optimal linear approximation to the eigenfunctions of the Laplace Beltrami operator on the manifold, which is capable of discovering nonlinear manifold structure. It uses the notion of the Laplacian of the graph to compute a transformation matrix which maps the data points to a subspace. Formally, it first constructs an undirected graph $G = (V, E)$, where each edge is associated with a weight W_{ij}. The elements of the weight matrix can, for instance, be computed by means of the heat kernel [8]

$$W_{ij} = \exp(-\sigma_w^{-2} \|x_i - x_j\|^2), \tag{8}$$

where σ_w is the width of the kernel. Based on the weight matrix W, it computes the graph Laplacian as $L = D - W$, where D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. To obtain the embeddings, the relationship between latent and observed variables is modeled as $z = \mathbf{F}x$, where $\mathbf{F}$ is the projection matrix. By imposing the orthonormal constraints ($xDx^T = I$), $\mathbf{F}$ is found in a closed form as a solution to the generalized eigenvalue problem:

$$x^T L x \mathbf{F} = \lambda x^T D x \mathbf{F}, \tag{9}$$

where the column vectors $\mathbf{F}_i$, $i = 1, ..., D_z$, with D_z being the dimension of the manifold, are the eigenvectors corresponding to the minimum eigenvalue solutions (λ_i) of (9). Thus, the projection $\mathbf{F}$ defines the manifold on which inputs $\mathbf{x}$ vary more smoothly.

3.2. Supervised Ordinal LPP (SO-LPP)

To obtain a manifold that is better adjusted to emotion classification, [19] proposed a Supervised Locality Preserving Projections (S-LPP) algorithm, which also encodes class information when computing the weight matrix in (8). We extend this algorithm by also encoding the ordering of the class labels, so as to preserve the smooth transitions between different emotion intensities on the manifold. We call this algorithm Supervised Ordinal LPP (SO-LPP) since its proximity matrix W is defined as an $ordinal$ weight matrix

W^{or}, with elements $(i,j) \in \{1 \dots N\}$ computed as

$$W_{ij} + \beta\, W_{\max} \sum_{k=1}^{R} I(h_i > k)\, I(h_j > k)\, I(c_i, c_j), \quad (10)$$

where W_{ij} is given in (8), $W_{\max} = \max_{\forall ij} W_{ij}$, β is the parameter that quantifies the degree of supervised learning, and $I(\cdot)$ is the indicator function defined in Sec. 2.1. In contrast to the similarity measure in (8), the similarity measure in (10) is augmented by the label information, thus increasing similarity *between* the samples belonging to the same emotion class and similarity *within* samples of the same emotion class based on their intensity levels. Note also that samples from different emotion classes, but with 'neutral' intensity will be all grouped together, where the samples with higher intensities will be 'shifted away' by the factor $\beta\, W_{\max}$ - which is exactly what we need for emotion classification and modeling of ordinal relationships between their intensities.

In the unsupervised setting, i.e. when the intensity levels are treated as latent variables, the elements of the ordinal weight matrix W^{or}_{ij} have the same form as in the supervised setting (10), with the only difference being that the indicator functions are replaced by the model estimates of the intensity levels $h \in \{1, ..., R\}$. Accordingly, the elements of this ordinal weight matrix W^{or}_{ij} have the following form:

$$W_{i,j} + \beta\, W_{\max} \sum_{k=1}^{R} p(h_i > k)\, p(h_j > k)\, I(c_i, c_j) \quad (11)$$
$$p(h_l > k) = 1 - \sum\nolimits_{m=1}^{k} p(h_l = m), \quad l = i, j,$$

where the probability $p(h_l = m)$ for the input x_l is estimated as explained in Alg.1 in Sec.4.1. Once the ordinal weight matrix W^{or} is constructed, it is used to compute the graph Laplacian L and projection matrix $\mathbf{F}$.

4. Laplacian SM-CORF (LSM-CORF)

In this section, we incorporate topology of our input data $\mathbf{x}$ into the SM-CORF model by constraining its parameters to lie on the ordinal manifold. This is achieved by enforcing the latent variables $\mathbf{u} \equiv \boldsymbol{\phi}(\mathbf{x}) = \mathbf{F}x$ to be a Gaussian Markov Random Field (GMRF) w.r.t. graph L (see [26] for details). Based on the GMRF representation, we obtain a prior over the latent variables $\mathbf{U} = [u_1 u_2 ... u_N]$ as

$$p(\mathbf{U}) = \prod_{i=1}^{n} p(u_i) = \frac{1}{Z_u} \exp\left(-\frac{\alpha}{2} tr(\mathbf{U}L\mathbf{U}^T)\right) \quad (12)$$

where $Z_{\mathbf{U}}$ is a normalization constant and $\alpha > 0$ is a scale parameter. Furthermore, since $\mathbf{z} = \mathbf{F}x$, the prior in (12) can be used to obtain a prior over the projection matrix $\mathbf{F}$ as

$$p(\mathbf{F}|x) = \frac{1}{Z_{\mathbf{F}}} \exp\left(-\frac{\alpha}{2} \mathbf{F}xLx^T\mathbf{F}^T\right) \quad (13)$$

The role of this prior is to enforce smoothness constraints on the manifold in which we intend to model ordinal regression. Note that these constraints are different from temporal constraints imposed by dynamic features in the SM-CORF model, since the former aim at preserving the topology of our input data.

By using the prior in (12), the likelihood function of the SM-CORF model given by (6) and by assuming a Gaussian prior over the model parameters Ω, $P(\Omega) = \mathcal{N}(\Omega|0, \gamma I)$, we obtain the posterior distribution

$$P(\mathbf{F}, \Omega|c, h, x) \propto P(c, h|x, \mathbf{F}, \Omega) P(\mathbf{F}|x) P(\Omega). \quad (14)$$

If we use the maximum a posteriori (MAP) strategy to estimate the projection matrix $\mathbf{F}$ and the model parameters Ω, the topology of our input data will be seamlessly integrated into the model [26]. We call this model the Laplacian SM-CORF (LSM-CORF) model. The importance of the GMRF-based prior in (13) can best be seen in terms of the graphical structure of the resulting model. Namely, this prior introduces an additional graphical structure into the SM-CORF model. Specifically, the graphical structure of the SM-CORF model alone has the form of a chain representing the explicit dependencies *only* between the labels of the neighbouring nodes. On the other hand, the graphical structure of the GMRF is richer in the sense that it captures dependencies between the labels over the whole dataset.

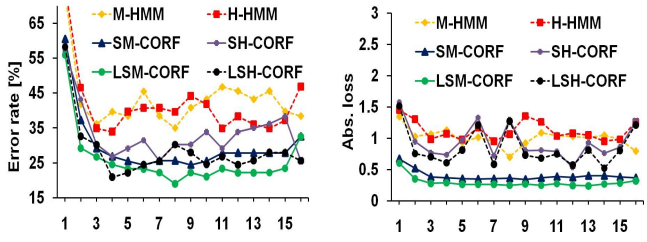
The MAP estimate of $(\mathbf{F}, \Omega)$ can be obtained by minimizing the following objective function:

$$\arg\min_{\mathbf{F},\Omega} \sum_{i=1}^{n} -\ln P(c_i, h_i|x_i, \mathbf{F}, \Omega) \quad (15)$$
$$+ \frac{\lambda}{2} \mathbf{F}XLX^T\mathbf{F}^T + \frac{\gamma}{2} \|\Omega\|^2 + const.$$

The penalty term $\mathbf{F}XLX^T\mathbf{F}^T$ has the role of manifold regularization in the LSM-CORF model, while λ and γ control the complexity of the projection matrix $\mathbf{F}$ and the ordinal regression model learned in the latent space, respectively. The Laplacian SH-CORF (LSH-CORF) model is obtained by replacing the likelihood term in (15) with (7) and by using the same analogy as before.

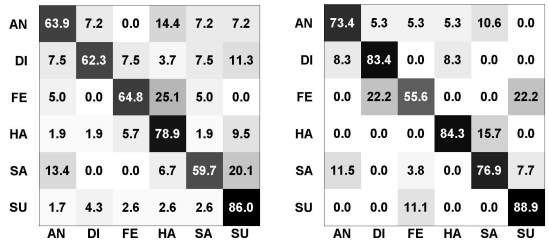
4.1. Model Learning

Parameter learning in the proposed model is performed by minimizing the objective function in (15) w.r.t. $(\mathbf{F}, \Omega)$ using the quasi-Newton limited-BFGS method (see [9] for the gradient derivation). In LSM-CORF, parameter learning is straightforward: first, we find an initial projection



(a) (b)

Figure 1. **BU-4DFE dataset.** The performance of the compared approaches w.r.t. the ordinal manifold dimensionality. (a) Mean error rates (in %) for facial expression recognition and (b) mean abs. loss for its intensity estimation.



(a) (b)

Figure 2. **BU-4DFE dataset.** Confusion matrices for facial expression recognition performed by (a) H-CORF and (b) LSH-CORF.

matrix $\mathbf{F}_0$ via the SO-LPP algorithm, and set the parameters Ω as in [9]. We then alternate between steps 1-2 in Alg.1 until convergence. Model learning in LSH-CORF requires some additional steps and this is described in Alg.1. The initial projection matrix $\mathbf{F}_0$ (i.e., the graph Laplacian) is obtained by dividing each training sequence $\mathbf{x_r}$ into R segments with approximately equal length, and by labeling each segment with the corresponding intensity level (i.e. the segment number). After one iteration of BFGS (step 2 in Alg.1), we use the *new* parameters $(\mathbf{F}, \Omega)$ to compute the likelihood of each intensity level h_i, where $i = 1, ..., R$. These likelihoods are then used to update graph Laplacian in (11). The steps 1-4 in Alg.1 are repeated until convergence of the evidence function.

Algorithm 1 Model Learning in LSH-CORF

Require: $\{c^i, h_0^i, \mathbf{x}^i\}_{i=1}^n$ and $(\mathbf{F}_0, \Omega_0)$

1. Evaluate the evidence in (15) and calculate the gradients w.r.t. $(\mathbf{F}, \Omega)$.
2. Feed the evidence and gradients to the BFGS method.
3. Calculate $P(h = i|\mathbf{F}, x, \Omega) = \sum_c P(c, h = i|\mathbf{F}, x, \Omega)$, where $i = 1, ..., R$.
4. Update graph Laplacian based on (11).
5. Repeat (1-4) until convergence of the evidence in (15).

5. Experiments

In this section we demonstrate the performance of the proposed method on the task of facial expression recognition and its intensity estimation from the frontal view facial images. We use image sequences from two publicly available datasets: the BU-4DFE dataset [24] and the Cohn-Kanade (CK) dataset [13]. Both datasets contain image sequences of different subjects displaying facial expressions of six prototypic emotions: Anger, Disgust, Fear, Happiness, Sadness and Surprise. We select 120 image sequences that come from 30 subjects from BU-4DFE, and 167 image sequences from 98 subjects from CK. All image sequences start with a neutral face evolving to the apex of the target display. Image sequences from the BU-4DFE dataset are sub-sampled so that the sequence lengths in both datasets are about 20-frame long on average. Each image sequence is annotated as one of six prototypic emotions ($c = \{1, ..., 6\}$), and each frame is manually labeled into three ordinal categories: neutral ($h = 1$) $\prec$ onset ($h = 2$) $\prec$ apex ($h = 3$).

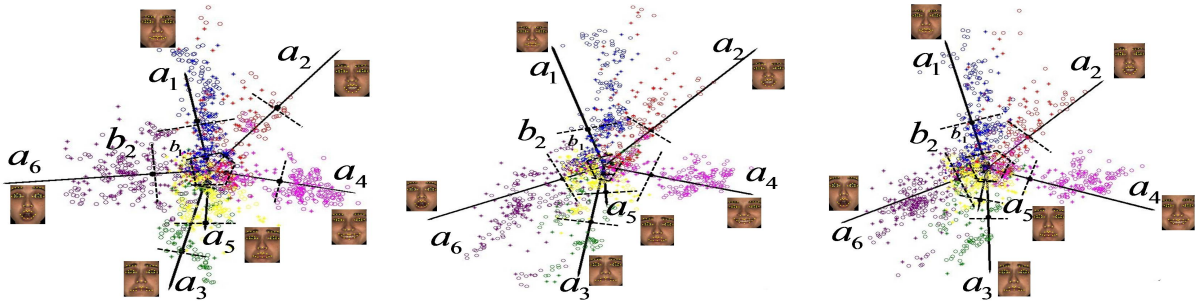
In this study, we use the locations of a set of characteristic facial points as the input features. In the case of BU-4DFE, we use 39 facial points extracted using the appearance based tracker [6]. For CK, we use 20 facial points extracted using the particle-filter-based tracker [15]. Fig. 6 depicts examples of the tracked sequences. The tracked points are later registered to a reference face and normalized w.r.t. the first frame in each image sequence. Finally, the PCA reduction preserving 95% of the total energy is applied to the input features, giving rise to the 16-dimensional inputs, for BU-4FE, and to the 24-dimensional inputs, for CK, which are denoted by $\mathbf{x}$.

We perform two sets of experiments. In the fully supervised setting, we compare the performance of our LSM-CORF model with: (1) Multi-output Hidden Markov Model (M-HMM), used as the baseline, (2) M-CORF and (3) SM-CORF. In the unsupervised setting, we perform the same experiments using the 'hidden' models (H-HMM/H-CORF/SH-CORF/LSH-CORF), all of which treat the intensity levels as latent variable. The M-HMM model is obtained by combining the outputs of standard HMM models trained independently for each emotion category using one-shot learning with h hidden states. In the unsupervised case (H-HMM), the initial estimates of the hidden states h are set as in LSH-CORF (Sec. 4.1). The emotion/level prediction for a given test sequence is accomplished using Viterbi decoding. Note that in this paper we do not include comparison with regular CRFs and *static* ordinal regression, since the state-of-the-art H-CORF [9] model has been shown to outperform those models in the target task.

In all our experiments, we apply 10-fold cross validation procedure, where each fold contains image sequences of different subjects. We report the accuracy using the mean

Table 1. **BU-4DFE dataset.** The performance of the compared approaches per emotion category. The ordinal manifold dimensionality which resulted in the best performance of the approach in question is used for the training/testing. Here we also include the results obtained by standard H/M-CORF models, which use an independent set of parameters for each emotion class, and are trained/tested using the original PCA-based feature vectors.

Method	Mean Error Rate for Facial Expression Recognition							Mean Absolute Loss for Facial Expression Intensity Prediction						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Ave.	Angry	Disgust	Fear	Happy	Sad	Surprise	Ave.
M-HMM	27.0	51.4	48.6	29.2	53.1	17.5	34.0	0.74	0.67	0.95	0.34	1.15	0.27	0.69
M-CORF	33.3	33.3	55.5	16.6	38.5	5.26	26.0	1.06	0.58	1.33	0.27	1.00	0.21	0.74
SM-CORF	58.3	15.8	44.4	11.1	30.7	6.67	24.0	1.17	0.32	1.00	0.28	0.92	0.27	0.66
LSM-CORF	31.6	15.7	33.3	5.55	26.1	0.00	**19.0**	0.75	0.21	0.66	0.11	0.46	0.00	**0.36**
H-HMM	27.0	40.3	51.4	28.1	60.8	12.5	36.7	1.00	0.90	1.40	0.76	2.09	0.51	1.11
H-CORF	36.1	37.7	35.2	21.1	40.3	14.0	30.1	1.2	0.79	1.40	0.45	1.6	0.35	0.96
SH-CORF	40.0	41.6	33.3	15.7	30.7	5.55	27.8	1.2	0.75	0.77	0.26	0.84	0.16	0.66
LSH-CORF	26.6	16.6	44.4	15.7	23.1	11.1	**22.9**	0.81	0.11	1.06	0.21	0.64	0.22	**0.50**



(a) it=80, 20.9% (0.14) (b) it=40, 21.5% (0.14) (c) it=80, 13.0% (0.12)

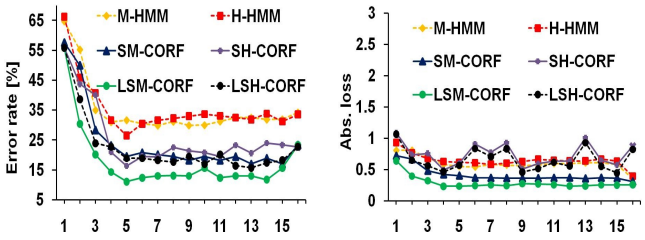
Figure 3. **BU-4DFE dataset.** Facial expression recognition and its intensity estimation achieved by (a) SM-CORF and (b-c) LSM-CORF, in the 3D ordinal manifold learned by the proposed OS-LPP. In SM-CORF, the embeddings remain unchanged during optimization of $\mathbf{\Omega} = \{\mathbf{a}_c, \mathbf{b}, \sigma, \mathbf{u}_c\}$, while in LSM-CORF, $\mathbf{\Omega}$ and the embedding matrix $\mathbf{F}$ are jointly optimized. Both algorithms converged in less than 80 iterations. Below each image, the error rates for facial expression recognition (in %) and mean abs. loss for the intensity estimation (obtained after the number of iterations (it)) are shown. Different colors in the images depict the embeddings of facial expressions of different emotion categories, and $(\cdot, *, \circ)$ correspond to their intensity levels (i.e., neutral, onset and apex), respectively.

error rate ($\frac{1}{N}\sum_n I(c_n \neq \bar{c}_n)$) for facial expression recognition, and mean absolute loss ($\frac{1}{NT}\sum_n \sum_t |h_{nt} - \bar{h}_{nt}|$) for its intensity estimation. Here, (c_n, h_{nt}) and $(\bar{c}_n, \bar{h}_{nt})$ are predicted and ground-truth emotion/intensity labels, respectively. The width of the heat kernel in (8) is set to the mean squared distance between the training inputs, and $\beta = 2$.

5.1. Experiments on the BU-4DFE dataset

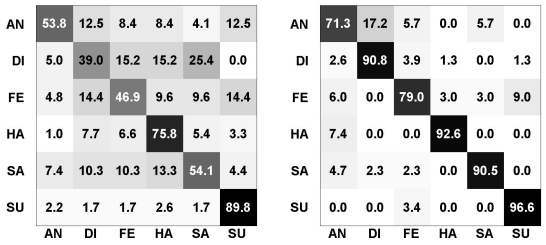
To select an optimal manifold for ordinal regression, we test the performance of the compared approaches w.r.t. the size of the ordinal manifold obtained as explained in Sec. 3. The average test errors for facial expression recognition and its intensity estimation are shown in Fig. 1. Here we only report results for SH/SM-CORF models since the regular H/M-CORF models were prone to severe overfitting on the manifold data. As can be seen from Fig. 1, all CORF-based models exhibit superior performance compared to that of H/M-HMM, with the proposed LSH/LSM-CORF performing the best. Table. 1 shows the performance of the tested models per each emotion category, trained/tested using optimal dimensionality of the ordinal manifold. The proposed approach outperforms other approaches in the tasks of facial expression recognition and its intensity estimation. The SH/SM-CORF models exhibit superior performance to that attained by standard H/M-CORF models, which can be attributed to their (1) effective parameter sharing and (2) modeling on the non-linear manifold specifically built for ordinal regression. However, the SH/SM-CORF models fail to further 'adapt' the ordinal manifold for modeling of dynamic ordinal regression. This is accomplished in LSH/LSM-CORF, leading to more accurate predictions. Confusion matrices for the H-CORF model [9] and the proposed LSH-CORF model are given in Fig. 2. The latter leads to better performance in all cases but the Fear class. A plausible explanation is that examples of Fear in BU-4FE often contain Smiles (of embarrassment) and acted Screams which are sources of confusion with Happiness and Surprise.

We also observed the manifold learning during the model estimation phase. For visualization purpose, we model the ordinal manifolds in 3D. Fig. 3 depicts adaptation of the LSM/SM-CORF models to the corresponding manifolds. Fig. 3(a) shows the SM-CORF model estimated on the 'fixed' manifold, while Fig. 3(b)-(c) show how this manifold changes during estimation of the proposed LSM-CORF model, which jointly estimates the manifold and the CORF



(a) (b)

Figure 4. **CK dataset.** The performance of the compared approaches w.r.t. the ordinal manifold dimensionality. (a) Mean error rates (in %) for facial expression recognition, and (b) mean abs. loss for its intensity estimation.



(a) (b)

Figure 5. **CK dataset.** Confusion matrices for facial expression recognition performed by (a) H-CORF and (b) LSH-CORF.

parameters. As can be seen from Fig. 3(a), the SM-CORF model is unable to handle overlap in examples of Disgust (a_2) and Happiness (a_4), since it uses *linear* projections a for each emotion class. On the other hand, the proposed LSM-CORF model handles this by simultaneously refining the ordinal manifold and estimating the ordinal regression parameters. Fig. 3 indicates parameter sharing among different CORF components (due to the similarity of 'neutral' and 'onset' of target emotions), which, in turn, leads to having a more discriminative model than is the case with the regular M-CORF model.

5.2. Experiments on the Cohn-Kanade dataset

Fig.4 shows the performance of the compared approaches w.r.t. the size of the ordinal manifold, while Table 2 shows the performance per emotion category obtained by using optimal ordinal manifolds to train/test the methods. The LSM-CORF model consistently outperforms other models, both in supervised and unsupervised setting. Interestingly, the proposed LSH-CORF model still accurately predicts emotion intensities, which is, in part, contributed to its modeling of the data topology. The confusion matrices in Fig. 5 similarly reflect superior performance of our LSH-CORF model compared to H-CORF [9], which we found to be prone to data overfitting.

5.3. Experiments on spontaneous data

We also test the applicability of the proposed approach on naturalistic data. To this end, we recorded a person while watching a funny video. We tracked the video obtained using the both trackers (i.e., [6, 15]), as in the experiments above. We then trained two separate LSM-CORF models using data from BU-4FE and CK. Fig.6 shows the tracking results as well as the quantitative results for continuous recognition of facial expressions of various emotions and their intensity estimation. Note that both models discriminate well between different emotions and give smooth predictions of their intensity levels. However, although both models classify the test sequence as a joyful display overall, the model trained on BU-4FE encodes high levels of Disgust. As can be seen from the bottom row in Fig. 6, which depicts the imagery from BU-4FE most similar to that tested, expressions similar to those depicted in the test video were labeled as Disgust in this dataset. On the other hand, the model trained on CK encodes Surprise in addition to Happiness, which is in agreement with manual annotation of the test video that we obtained by asking three lay experts to score the video in terms of six basic emotion categories.

6. Conclusions

Modeling the intrinsic topology of the facial affect data is important for educing discriminative features for dynamic recognition of emotions and their intensity. Standard generative models like HMMs and discriminative models like H-CORF [9] use simple *linear* feature representation that is unable to capture such topology. In contrast, the proposed LSM-CORF model incorporates this topology into the H-CORF framework, giving rise to a linear approximation of the otherwise non-linear model for dynamic ordinal regression. As evidenced by the results, the proposed method attains effective simultaneous dynamic recognition and intensity estimation of multiple emotions on both deliberately and spontaneously displayed facial expressions.

Acknowledgments

We are grateful to Minyoung Kim for his help throughout the course of this work. This material is based upon work supported by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB), and by the National Science Foundation under Grant No. IIS 0916812.

References

[1] Z. Ambadar, J. Schooler, and J. Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 16(5):403–410, 2005. 1

[2] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin*, 111:256–274, 1992. 1

Table 2. **CK dataset.** The performance of the compared approaches per emotion category. The ordinal manifold dimensionality which resulted in the best performance of the approach in question is used for the training/testing.

Method	Mean Error Rate for Facial Expression Recognition							Mean Absolute Loss for Facial Expression Intensity Prediction						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Ave.	Angry	Disgust	Fear	Happy	Sad	Surprise	Ave.
M-HMM	50.7	22.0	35.0	15.6	49.8	9.70	30.5	0.68	0.36	0.48	0.28	1.19	0.05	0.50
M-CORF	38.7	32.0	20.1	20.5	33.3	8.57	25.5	1.25	0.68	0.48	0.54	0.76	0.17	0.64
SM-CORF	35.5	16.1	24.0	2.70	42.8	2.85	20.9	0.81	0.32	0.40	0.05	1.00	0.14	0.45
LSM-CORF	23.0	12.1	8.00	2.70	23.8	2.85	**12.0**	0.75	0.16	0.16	0.05	0.67	0.14	**0.32**
H-HMM	60.0	22.0	22.0	12.7	48.1	15.7	35.8	1.06	0.16	0.20	0.05	1.00	0.20	0.45
H-CORF	46.2	61.0	53.1	24.2	45.9	10.2	40.0	1.18	1.28	0.56	0.44	0.52	0.34	0.72
SH-CORF	32.5	8.00	22.0	2.72	32.3	2.85	16.7	1.12	0.12	0.44	0.05	1.47	0.11	0.55
LSH-CORF	28.7	9.20	21.0	7.40	9.50	3.40	**13.2**	1.06	0.24	0.52	0.17	0.28	0.08	**0.39**

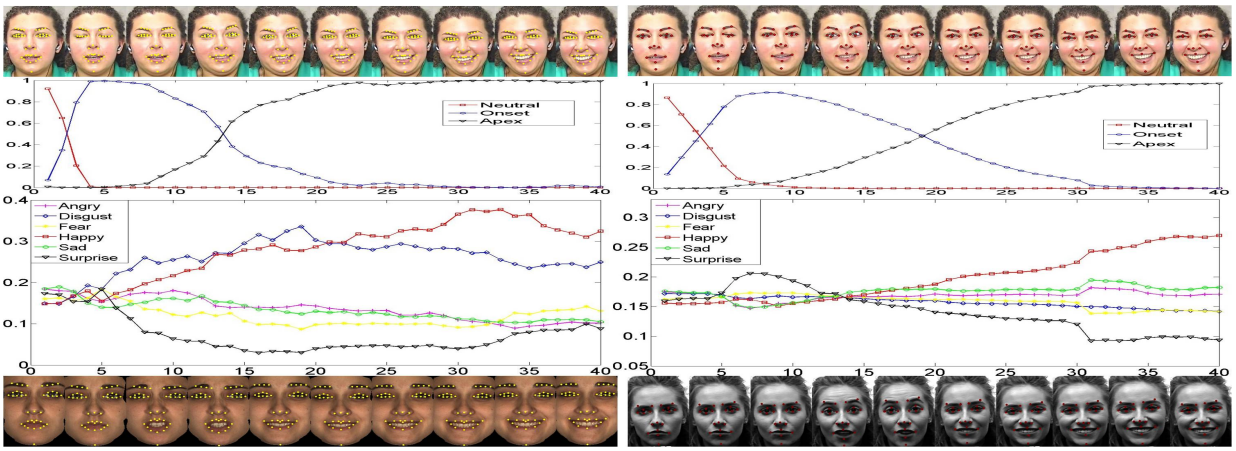


Figure 6. **Continuous prediction of naturalistic expression of emotion.** The shown images are subsampled from the test sequence by factor four. The graphs in between show the estimated probabilities for various facial expressions of emotions and their intensities obtained by the proposed LSM-CORF model. The models are trained using the data from BU-4FE ($left$), and CK ($right$), some examples of which are shown in the bottom row.

[3] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *JMLR*, 6:1019–1041, 2005. 2

[4] W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. pages 145–152, 2005. ICML. 2

[5] J. Delannoy and J. McDonald. Automatic estimation of the dynamics of facial expression using a three-level model of intensity. *FG*, pages 1–6, 2008. 1

[6] F. Dornaika and J. Orozco. Real time 3d face and facial feature tracking. *J. Real-Time Image Processing*, pages 35–44, 2007. 5, 7

[7] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Trans. on Systems, Man, and Cybernetics*, 39(1):64–84, 2009. 1

[8] X. He and P. Niyogi. Locality Preserving Projections. *NIPS*, 2004. 3

[9] M. Kim and V. Pavlovic. Hidden conditional ordinal random fields for sequence classification. *Machine Learning and Knowledge Discovery in Databases*, 6322:51–65, 2010. 1, 2, 3, 4, 5, 6, 7

[10] M. Kim and V. Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. *ECCV*, pages 649–662, 2010. 1, 2

[11] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE PAMI*, 32:1940–1954, 2010. 1

[12] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 2001. ICML. 2

[13] J. Lien, T. Kanade, J. Cohn, and C. Li. Detection, tracking, and classification of action units in facial expression. *J. of Rob. and Aut.Systems*, 31(3):131–146. 5

[14] M. Mahoor, S. Cadavid, D. Messinger, and J. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. *CVPRW*, pages 74–80, 2009. 1

[15] I. Patras and M. Pantic. Particle Filtering with Factorized Likelihoods for Tracking Facial Features. *FG*, pages 97–102, 2004. 5, 7

[16] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. *NIPS*, pages 1097–1104, 2004. 3

[17] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recognition*, 25(1):65–77, 1992. 1

[18] A. Savrana, B. Sankur, and M. Bilgeb. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 2012. 1

[19] C. Shan, S. Gong, and P. W. Mcowan. Appearance manifold of facial expression. *Lecture Notes in Comp. Science*, 3766:221–230, 2005. 3

[20] C. Shan, S. Gong, and P. W. Mcowan. Dynamic facial expression recognition using a bayesian temporal manifold model. *BMVC*, pages 297–306, 2006. 1

[21] A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. pages 973–944, 2002. NIPS. 2

[22] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Trans. on Systems, Man, and Cybernetics*, 42(1):28–43, 2012. 1

[23] P. Yang, Q. Liu, and D. N. Metaxas. Rankboost with l1 regularization for facial expression recognition and intensity estimation. *ICCV*, pages 1018–1025, 2009. 1

[24] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. *FG*, pages 679–684, 2008. 5

[25] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE PAMI*, 31:39–58, 2009. 1

[26] G. Zhong, W. Li, D. Yeung, X. Hou, and C. Liu. Gaussian process latent random field. 2010. AAAI. 4