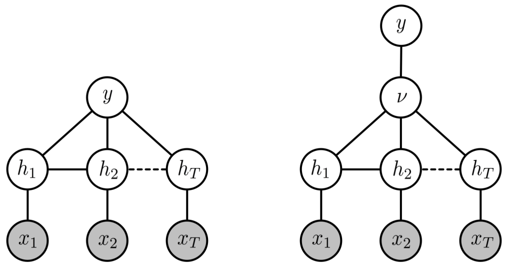
Variable-state Latent Conditional Random Fields for Facial Expression Recognition and Action Unit Detection

Robert Walecki[1], Ognjen Rudovic[1], Vladimir Pavlovic[2] and Maja Pantic[1]
[1] Computing Department, Imperial College London, UK
[2] Department of Computer Science, Rutgers University, USA

Abstract— **Automatic recognition of facial expressions of emotions, and detection of facial action units (AUs), from videos depends critically on modeling of their dynamics. These dynamics are characterized by changes in temporal phases (onset-apex-offset) and intensity of emotion/AUs, the appearance of which vary considerably among subjects, making the recognition/detection task very challenging. While state-of-the-art Latent Conditional Random Fields (LCRF) allow one to efficiently encode these dynamics via modeling of structural information (e.g., temporal consistency and ordinal constraints), their latent states are restricted to either unordered (nominal) or fully ordered (ordinal). However, such an approach is often too restrictive since, for instance, in the case of AU detection, the sequences of an active AU may better be described using ordinal latent states (corresponding to the AU intensity levels), while the sequences of this AU not being active may better be described using unordered (nominal) latent states. To this end, we propose the Variable-state LCRF model that automatically selects the optimal latent states (nominal or ordinal) for each sequence from each target class. This unsupervised adaptation of the model to individual sequence or subject contexts opens the possibility for improved model fitting and, subsequently, enhanced predictive performance. Our experiments on four public expression databases (CK+, AFEW, MMI and GEMEP-FERA) show that the proposed model consistently outperforms the state-of-the-art methods for both facial expression recognition and action unit detection from image sequences.**



(a) H-CRF[5]/H-CORF[6]

(b) VSL-CRF

Fig. 1: The graph structure of the (a) traditional Latent CRF models H-CRF/H-CORF, and (b) proposed VSL-CRF model. In H-CRF/H-CORF, the latent states h, relating the observation sequence $\mathbf{x} = \{x_1, \dots, x_T\}$ to the target label y (e.g., emotion or AU activation). are allowed to be either nominal or ordinal, while in VSL-CRF the latent variable $\nu = \{nominal, ordinal\}$ performs automatic selection of the optimal latent states for each sequence from each class.

I. Introduction

Facial behavior is believed to be the most important source of information when it comes to affect, attitude, intentions, and social signals interpretation. Automatic facial expression recognition has, therefore, been an active research topic for more than two decades [1]. Facial expressions are typically described at two levels: the facial affect (emotion) and facial muscle actions (AUs), which stem directly from the message and sign judgment approaches for facial expression measurement [2]. The message judgment aims to directly decode the meaning conveyed by a facial display (e.g., in terms of the six basic emotions), while the sign judgment instead aims to study the physical signal used to transmit the message (such as raised cheeks or depressed lips). The *Facial Action Coding System* (FACS) [3] is the most comprehensive, anatomically-based system for encoding expression by describing facial activity on the basis of 33 unique AUs, as well as several categories of head and eye positions and other movements [4].

Early research on facial expression analysis focused mainly on recognition of prototypic facial expressions of six basic emotions (anger, happiness, fear, surprise, sadness, and disgust) and detection of AUs from static facial images [1]. However, recognizing facial expressions from videos (i.e., image sequences) is more natural and has proved to be more effective [7]. These is motivated by the fact that facial expressions can better be described as a dynamic process that evolves over time. For instance, the facial expression of Happiness is usually characterized by its temporal phases (onset-apex-offset). Similarly, the activation of AUs spans different time intervals that reflect variation in their temporal phases and intensity (which can also be described using FACS).

Most state-of-the-art approaches for facial expression analysis [6, 7, 8, 9] focus on modeling of the spatio-temporal dynamics of facial expressions in order to improve their recognition. These methods can be cast as variants of the class of conditional models called Latent Conditional Random Fields (LCRF) [5], which have also been applied successfully in other domains (e.g. gesture recognition [5] and human motion estimation [10]) to encode the dynamics of the target tasks. In the context of facial expressions, LCRFs have been used to model temporal dynamics of facial expressions as a sequence of latent states, relating the image features to the class label (e.g., the emotion category). A typical representative of these models is the Hidden CRF (H-CRF) [5, 11, 12, 13], used for facial expression recognition of six basic emotions. However, apart from temporal constraints imposed on its latent states, this model does not assume any other structure in the data. On the other hand, the recently proposed Hidden Conditional Ordinal Random Field (H-CORF) model [6, 8] imposes additional constraints on the latent states of emotions by exploiting their ordinal relationships. This model attempts to correlate the latent states of emotions with their temporal phases (or intensity) by representing them on an ordinal scale. This, in turn, results in the model with fewer parameters that is able to discriminate better between facial expressions of different emotions.

However, in the LCRF models such as H-CRF and H-CORF, and their variants, the latent states are assumed to be either nominal or ordinal for each and every class. This representation can be too restrictive since for some classes modeling the latent states as ordinal may help to better capture the structure of the states, i.e., their ordinal relationships, allowing the model to better fit the data. By contrast, it would be wrong to impose ordinal constraints on latent states of the classes that do not exhibit ordinal structure. In this case, the more flexible nominal model will better fit the data. For example, we expect the latent states of the emotion class Happiness to be correlated with its temporal phases defined on an ordinal scale (neutral<onset<apex). On the other hand, facial expressions of neutral or a mix of sequence of other emotions are not expected to have ordered latent states, so fitting nominal model is a more natural choice. Similarly, for the task of AU detection, attempting to represent its temporal phases or intensity levels via ordinal states seems natural; however, modeling the latent states of the negative class (all other sequences not containing the target AU) is expected to be more effective when nominal states are used due to the lack of the ordinal structure as well as high variability in such data. The duality of nominal and ordinal representation can exist even within a single emotion or AU class. This can occur due to the difference in the facial expressiveness of different subjects, or due to the clustering effects of the features caused by the subject-specific variation dominating that related to the facial expressions. Thus, the model should be able to automatically select which type of the latent states is optimal for modeling individual sequences from each of target classes.

In this paper, we generalize the LCRF models by relaxing their assumption that the latent states within target classes need only be nominal or ordinal. We do so by allowing the model to use both types of latent states for modeling sequences within and across the classes, in order to improve their classification. To this end, we introduce the Variable-state LCRF (VSL-CRF) model that can automatically select the optimal model family for each sequence from each class. This is achieved by means of the newly introduced latent variable ν in the graph structure of LCRFs, which performs selection of the optimal feature functions in the model (nominal or ordinal) on the sequence level (see Fig.1). The selection of the latent states is attained in a fully unsupervised manner via a max-pooling of the nominal/ordinal node potentials during learning and inference in the VSL-CRF model. In contrast to H-CRF/H-CORF models, this makes the objective function of our model non-smooth, thus the standard gradient-based optimization methods for parameter learning can not directly be applied. For this, we propose a learning approach based on the notion of the function sub-gradients [14], which allows us to efficiently learn the model parameters. We show on four publicly available datasets that the proposed VSL-CRF model outperforms existing LCRF-based models, and other state-of-the-art models for the target tasks.

II. Related Work

In this section, we briefly review the most recent works on facial expression recognition and facial action unit detection.

A. Facial Expression Recognition

Facial expression recognition methods can be categorized into frame-based and sequence-based (see [15] for a detailed overview). Frame-based methods attempt the expression recognition from a single image (typically, the apex of the expression) [16, 17, 18]. However, a natural facial event is dynamic, which evolves over time from the onset, the apex, to the offset. Therefore, recognizing facial expressions from videos is more natural. Although some of the frame-based methods use the features extracted from several frames in order to encode dynamics of facial expressions, models for dynamic classification provide a more principled way of doing so. With a few exceptions, most of the dynamic approaches to classification of facial expressions are based on variants of Dynamic Bayesian Networks (DBN) (e.g., Hidden Markov Models (HMM) [19] and Conditional Random Fields (CRF) [20]). For example, [21] trained independent HMMs for each emotion category, and then performed emotion classification by comparing the likelihoods of the HMMs. However, discriminative models based on CRFs have been shown to be more effective for the expression classification [12, 13, 22]. For instance, [12, 22] used a generalization of the linear-chain CRF model, a Hidden Conditional Random Field (H-CRF) [5], where additional layer of (hidden) variables is used to model temporal dynamics of facial expressions. The training of the model was performed using image sequences, but classification of the expressions was done by selecting the most likely class (i.e., emotion category) at each time instance. The authors showed that: (i) having the additional layer of hidden variables results in the model being more discriminative than the standard linear-chain CRF, and (ii) that modeling of the temporal unfolding of the facial shapes is more important than their spatial variation for discriminating between different facial expressions (based on comparisons with SVMs). Another modification of H-CRF, named partially-observed H-CRF, was proposed in [13], where additional hidden variables are added to the model to encode the occurrence of subsets of AU combinations in each image frame, and which are assumed to be known during learning. This method outperformed the standard H-CRF, which does not use a prior information about the AU co-currencies. In contrast to these models, [6, 8] proposed the Hidden Conditional Ordinal Random Field (H-CORF) models, which encode ordinal relationships between the temporal phases of emotion. These models outperformed the nominal H-CRF models, which fail to impose the order constraints on their latent states. Nevertheless, the main limitation of the models mentioned above is that they restrict their latent states to be either nominal or ordinal, but not both.

B. Facial AU Detection

As for facial expression recognition, two main approaches have been proposed for AU detection: static and temporal modeling. In the former, image features are extracted from each frame and then fed into a static classifiers such as SVM or AdaBoost [23]. More recently, [4] proposed a method based on static SVMs, named Selective Transfer Machine (STM), which personalizes the generic SVM classifier by learning the classifier and re-weighting the training samples that are most relevant to the test subject during inference. The resulting method achieves the state-of-the-art results on the AU detection task, considerably outperforming generic SVMs. However, this comes at the cost of transductive learning of this approach. Temporal modeling for AU detection has been attempted using either temporal image features [24, 25] or DBN-based models such as HMMs [26] and CRFs [27]. All these methods perform the frame-based AU detection despite the fact that an AU activation typically ranges over several or more image frames. To the best of our knowledge, the works that attempted the sequence-based AU detection perform either majority voting using the frame-based detection [23], or detection of the temporal phases of AUs followed by the rule-based classification of the sequences (by detecting the onset-apex-offset sequence of an AU) [26, 28]. Other temporal models based on Ordinal CRFs have been proposed for modeling of AU temporal phases [29], and their intensity [9], however, they do not perform AU detection.

Common to the approaches for facial expression recognition and AU detection mentioned above is that they all use either static/dynamic classifiers which are designed for either nominal or ordinal data. While the former imposes no spatial constraints on target classes, the latter does so for all classes (e.g., all emotions are modeled by imposing ordinal constraints). In the context of the temporal models based on CRFs, this results in the models that are either underconstrained (e.g., H-CRF[5]) or overconstrained (H-CORF[6]), which limits their representational power. To mitigate this, the proposed VSL-CRF model allows different classes to be of the nominal or ordinal type, which is inferred from target data. In what follows, we first introduce the proposed method. We then show our experimental evaluation and conclude the paper.

III. Methodology

We consider a K-class classification problem, where we let $y \in \{1, ..., K\}$ be the class label (e.g., emotion category). Each class y is further represented with a sequence of (latent) states denoted as consecutive integers $h \in \{1, \dots, C\}$, where C is the number of possible states (e.g., temporal phases such as neutral-onset-apex of emotion). The sequence of the corresponding image features, denoted by $\mathbf{x} = \{x_1 \dots x_T\} \in T \times D$, serves as input covariates for predicting y and $\mathbf{h} = (h_1, \dots, h_T)$. The length of sequences T can vary from instance to instance, while the input feature dimension D is constant. If not said otherwise, we assume a supervised setting where 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.

A. Conditional Random Fields (CRF)

CRFs [30] are a class of log-linear models that represent the conditional distribution $P(\mathbf{h}|\mathbf{x})$ as the Gibbs form clamped on the observation $\mathbf{x}$:

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

Here, $Z(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{h} \in \mathcal{H}} e^{s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta})}$ is the normalizing partition function ($\mathcal{H}$ is a set of all possible output configurations), and $\boldsymbol{\theta}$ are the parameters[1] of the *score function* (or the negative energy) $s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta})$. Note that in this model, the states $\mathbf{h}$ are observed and they represent the frame labels.

We further assume the linear-chain graph structure $G = (V, E)$ in the model, described by the *node* ($r \in V$) and *edge* ($e = (r, s) \in E$) potentials. We denote the node features by $\mathbf{\Psi}_r^{(V)}(\mathbf{x}, h_r)$ and the edge features by $\mathbf{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s)$. By letting $\boldsymbol{\theta} = \{\mathbf{v}, \mathbf{u}\}$ be the parameters of the node and edge potentials, respectively, $s(\mathbf{x}, \mathbf{h}; \boldsymbol{\theta})$ can then be written as the sum:

$$\sum_{r \in V} \mathbf{v}^\top \mathbf{\Psi}_r^{(V)}(\mathbf{x}, h_r) + \sum_{e=(r,s) \in E} \mathbf{u}^\top \mathbf{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s). \tag{2}$$

Although the representation in (2) is so general that it can subsume nearly arbitrary forms of features, the node/edge features are often defined depending on target task. We limit our consideration to two commonly used types of the node features (nominal/ordinal), which can be represented using a general probabilistic model for static modeling of nominal/ordinal classes. This is achieved by setting the potential at node r as $\mathbf{v}^\top \mathbf{\Psi}_r^{(V)}(\mathbf{x}, h_r) \longrightarrow \mathbf{\Gamma}_r^{(V)}(\mathbf{x}, h_r)$, where

$$\Gamma_r^{(V)}(x, h_r) = \sum_{c=1}^{C} \mathrm{I}(h_r = c) \cdot logP(h_r = c|f(x)). \tag{3}$$

The **nominal** node potential is then obtained by using the multinomial logistic regression (MLR) model [5]:

$$P(h_r^n = c|f^n(x, c)) = \frac{\exp(f^n(x, c))}{\sum_{l=1}^{C} \exp(f^n(x, l))}, \tag{4}$$

where $f_n(x, c) = \beta_c^T \cdot [1, x]$, for $c = 1, ..., C$, and β_c is the separating hyperplane for the c-th *nominal* state of the target class. By plugging the likelihood function in (4) into

[1] For simplicity, we often drop the dependency on $\boldsymbol{\theta}$ in notations.

the node potential in (3), we obtain the node features of the standard CRF model.

Recently, several authors proposed using the ranking likelihood to define the **ordinal** node potentials. This likelihood is derived from the threshold model for (static) ordinal regression [31], and has the form:

$$P(h_r^o = c|f^o(x,c)) = \Phi(\frac{b_c - f^o(x)}{\sigma}) - \Phi(\frac{b_{c-1} - f^o(x)}{\sigma}), \tag{5}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function (c.d.f.), and $f^o(x) = a^T(x)$. The parameter vector a is used to project the input features onto an *ordinal* line divided by the model thresholds or cut-off points $b_0 = -\infty \leq \cdots \leq b_C = \infty$, with each bin corresponding to one of the *ordinal* states $c = 1, ..., C$ in the model. The ranking likelihood in (5) is constructed by contaminating the ideal model (see [32] for details) with Gaussian noise with standard deviation σ. Again, by plugging the likelihood function in (5) into the node potential in (3), we obtain the node features of the Ordinal CRF (CORF) model [32].

In both models defined above (the standard CRF and CORF), the edge potentials $\mathbf{\Psi}_e^{(E)}(\mathbf{x}, h_r, h_s)$ are defined in the same way and have the form:

$$\Big[I(h_r = c \ \wedge \ h_s = l)\Big]_{C\times C} \times \big|x_r - x_s\big|, \tag{6}$$

where $I(\cdot)$ is the indicator function that returns 1 (0) if the argument is true (false). The role of the edge potentials is to assure the temporal consistency of the nominal/ordinal states within a sequence.

B. Latent Conditional Random Fields (LCRFs)

While the CRFs introduced in the previous section aim at modeling/decoding of the state-sequence within a single class, the framework of LCRFs [5, 33] aims at the sequence level multi-class classification. This is attained by introducing additional node in the graph structure of CRF/CORFs (see Fig.1) representing the class label, where the latent states $\mathbf{h}$ are now treated as unknown. Formally, LCRFs combine the score functions of K CRFs, one for each class $y = \{1, \ldots, K\}$, within the following score function:

$$s(y, \mathbf{x}, \mathbf{h}; \mathbf{\Omega}) = \sum_{k=1}^{K} I(k = y) \cdot s(\mathbf{x}, \mathbf{h}; \theta_y), \tag{7}$$

where $s(\mathbf{x}, \mathbf{h}; \theta_y)$ is the y-th CRF score function, defined as in (2), and $\mathbf{\Omega} = \{\theta_k\}_{k=1}^K$ denotes the model parameters. With such score function, the joint conditional distribution of the class and state-sequence is defined as:

$$P(y, \mathbf{h}|\mathbf{x}) = \frac{\exp(s(y, \mathbf{x}, \mathbf{h}))}{Z(\mathbf{x})}. \tag{8}$$

The sequence of the states $\mathbf{h} = (h_1, \ldots, h_T)$ is unknown, and they are integrated out by directly modeling the class conditional distribution:

$$P(y|\mathbf{x}) = \sum_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{x}) = \frac{\sum_{\mathbf{h}} \exp(s(y, \mathbf{x}, \mathbf{h}))}{Z(\mathbf{x})}. \tag{9}$$

Evaluation of the class-conditional $P(y|\mathbf{x})$ depends on the partition function $Z(\mathbf{x}) = \sum_k Z_k(\mathbf{x}) = \sum_k \sum_{\mathbf{h}} \exp(s(k, \mathbf{x}, \mathbf{h}))$, and the class-latent joint posteriors $P(k, h_r, h_s|\mathbf{x}) = P(h_r, h_s|\mathbf{x}, k) \cdot P(k|\mathbf{x})$. Both can be computed from independent consideration of K individual CRFs. The model with the *nominal* node potentials in the score function in (9) is termed Hidden CRF (H-CRF) [5]. Likewise, the model with the *ordinal* node potentials is termed Hidden CORF (H-CORF) [6].

C. Variable-state Latent Conditional Random Fields (VSL-CRF)

In this section, we generalize the H-CRF/H-CORF models by allowing their latent states to be modeled using either nominal or ordinal potentials within each class. In this way, we allow the model to select in an unsupervised manner the optimal feature functions for representing the target sequences. In what follows, we provide a formal definition of the model, and then explain its learning and inference.

VSL-CRF Model

Definition (Variable-state Latent CRF) *Let $\nu = (\nu_1, \ldots, \nu_K)$ be a vector of symbolic states or labels encoding the nature of the latent states h^ν of the i-th sequence, $i = 1, \ldots, N_y$ from class $y = (1, ..., K)$, either as nominal ($\nu_y = 0$) or ordinal ($\nu_y = 1$). The score function for class y in the VSL-CRF model is then defined as:*

$$s(y, \mathbf{x}, \mathbf{h}, \nu; \mathbf{\Omega}) = \begin{cases} \sum\limits_{k=1}^{K} I(k = y) \cdot s(\mathbf{x}, \mathbf{h}; \theta_y^n), & \text{if}\, \nu_y = 0 \\ \sum\limits_{k=1}^{K} I(k = y) \cdot s(\mathbf{x}, \mathbf{h}; \theta_y^o), & \text{if}\, \nu_y = 1 \end{cases} \tag{10}$$

where the nominal ($s(\mathbf{x}, \mathbf{h}; \theta_y^n)$) / ordinal ($s(\mathbf{x}, \mathbf{h}; \theta_y^o)$) score functions represent the sum of the node and edge potentials in (3) and (6), respectively. The class conditional distribution is then defined as:

$$P(y|\mathbf{x}) = \frac{\max\limits_{\nu}(\sum\limits_{h} \exp(s(y, \mathbf{x}, \mathbf{h}, \nu)))}{Z(\mathbf{x})}, \tag{11}$$

where $Z(\mathbf{x}) = \sum\limits_k Z_k(\mathbf{x}) = \sum\limits_k \max\limits_{\nu}(\sum\limits_h \exp(s(k, \mathbf{x}, \mathbf{h}, \nu)))$ and $\mathbf{\Omega} = \{\theta_k^n, \theta_k^o\}_{k=1}^K$.

We make a few remarks about the model defined above. The sequences from class y are modeled using either the nominal or ordinal latent states. The assignment of target sequences to nominal/ordinal models is therefore performed automatically, i.e., without need for any prior knowledge (supervision) about the underlying structure of the latent states capturing the dynamics of each sequence in the target class. Furthermore, since we use the max-pooling strategy in the class conditional distribution, we restrict the model to representing each sequence using either nominal or ordinal states. This prevents the redundancy in the representation of each sequence, which would otherwise occur if we performed the standard integration over the indicator variables

ν. However, in contrast to the objective functions of the H-CRF and H-CORF models, this results in the non-smooth objective function of the model.

VSL-CRF Learning and Inference

The parameter optimization in the H-CRF/H-CORF models is carried out by maximizing the (regularized) negative log-likelihood of the class conditional distribution in (9). Furthermore, to avoid the constrained optimization in H-CORF (due to the order constraints in parameters $\mathbf{b}$ of the ordinal node potentials), the displacement variables γ_c, where $b_j = b_1 + \sum_{k=1}^{j-1} \gamma_k^2$ for $j = 2, \ldots, C-1$ are introduced. So, $\mathbf{b}$ is replaced by the unconstrained parameters $\{b_1, \gamma_1, \ldots, \gamma_{C-2}\}$. Similarly, the positivity of the ordinal scale parameter is ensured by setting $\sigma = \sigma_0^2$. Although both the objectives of H-CRF/H-CORF are non-convex because of the logpartition function (log-sum-exp of nonlinear concave functions). However, their log-likelihood objective is bounded below by 0 and are both smooth functions, so standard quasi-Newton (such as Limited-memory BFGS) or the stochastic gradient descent algorithms can be used to estimate the model parameters (we use the former). Unfortunately, the objective function of the VSL-CRF model is both non-convex and non-smooth because of the *max* function in its class conditional distribution. Therefore, the gradients of the objective w.r.t. the (unconstrained) model parameters $\boldsymbol{\Omega}$ cannot be directly computed. Yet, the nominal/ordinal score functions are both subdifferentiable. We use this property to construct the subgradient [14] of the VSL-CRF objective at $\boldsymbol{\Omega} = \{\theta_k^n, \theta_k^o\}_{k=1}^K$. Formally, the VSL-CRF objective is given by:

$$RLL(\boldsymbol{\Omega}) = -\sum_{i=1}^{N} \log P(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\Omega}) + \lambda_{n(o)}||\theta_{k=1..K}^{n(o)}||^2, \tag{12}$$

where we introduce L-2 regulizers over the parameters of the nominal/ordinal score functions, the effect of which is controlled by λ_n/λ_o (set using a validation procedure).

The most critical factor that differentiates the minimization of the objective in (12) and that of the H-CRF/H-CORF models is the need to compute the subgradients $g \in \partial RLL(\boldsymbol{\Omega})$ of the objective function. Practically, this boils down to computing the following subgradients:

$$\nabla \max_{\nu}(\sum_{h} \exp(s(k, \mathbf{x}, \mathbf{h}, \nu))),\ k = 1, \ldots, K,$$

which are further given by

$$\begin{cases} \nabla \sum\limits_{h} \exp(s(\mathbf{x}, \mathbf{h}, \theta_k^n)), \\ \qquad\qquad if\ \sum\limits_{h} \exp(s(\mathbf{x}, \mathbf{h}, \theta_k^n)) > \sum\limits_{h} \exp(s(\mathbf{x}, \mathbf{h}, \theta_k^o)) \\ \nabla \sum\limits_{h} \exp(s(\mathbf{x}, \mathbf{h}, \theta_k^0)),\ otherwise. \end{cases}$$

Thus, at a point $\boldsymbol{\Omega}^*$ where one of the score functions, say nominal, gives a higher score than the ordinal for the given sequence, $\max\limits_{\nu}(\sum\limits_{h} \exp(s(k, \mathbf{x}, \mathbf{h}, \nu)))$ is differentiable and has the gradient $\partial\theta_k^n = \nabla \sum\limits_{h} \exp(s(\mathbf{x}, \mathbf{h}; \theta_k^n))$, while $\partial\theta_k^o = 0$. In other words, to find a subgradient of the maximum of the score functions, we choose the score functions that achieves the maximum for the target sequence at the current parameters, and compute the gradient of that score function only. Once this is performed, the gradient derivation is the same as in the H-CRF/H-CORF models (see [6] for more details). Finally, the assignment of a test sequence to the particular class, such as the facial action or emotion, is accomplished by the MAP rule $y^* = \mathrm{argmax}_y P(y|\mathbf{x}^*)$.

Method (6 basic emotions)	F-1 [%]	Accuracy [%]
Bartlett et al. [38]	—	87.5
PO-HCRF9 [13]	—	92.9
TMS [12]	—	91.2
SVM-SB	89.8	91.3
H-CRF	91.2	93.2
H-CORF	90.4	92.3
VSL-CRF	**94.5**	**96.7**

Method (7 emotions)	F-1 [%]	Accuracy [%]
ITBN [39]	—	86.3
STM-ExpLet [7]	—	94.2
SVM-SB	89.5	91.1
H-CRF	85.0	89.1
H-CORF	91.7	93.5
VSL-CRF	**93.9**	**95.8**

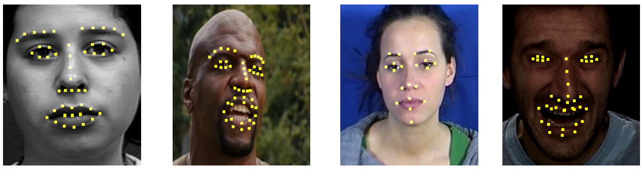
TABLE I: CK+ dataset. The upper part of the table shows the average results obtained on the 6 basic emotions, while the lower shows the results obtained on 6 basic emotions + contempt. We used these two sets of emotions to have direct comparisons with the state-of-the-art methods, which were evaluated on the same emotion sets.

IV. Experiments

We evaluated the proposed model on four publicly available facial expression datasets: Extended Cohn-Kanade (CK+) [34], AFEW Database [35], MMI Database [36] and GEMEP-FERA [37]. From CK+, we used 327 sequences labeled as Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise or Neutral. For this dataset, we performed 10-fold subject-independent cross-validation. The AFEW dataset has been collected from the movie videos showing close-to-real-world conditions, comprising video clips that have been labeled in terms of six basic emotions and neutral. To perform direct comparisons with the state-of-the-art method [7] for this dataset, we used the same experimental setting. For AU detection, we used the MMI dataset, where we chose five most frequent AUs from the upper face $(1, 2, 4, 6, 7)$. We created 4 folds for the training/test in a subject-independent manner. For each AU, the image sequences containing target AU were used as the positive class, while the remaining sequences were used as the negative class. We also used the GEMEP-FERA challenge dataset for AU detection. As in [4], we performed the leave-one-subject-out cross validation of the models for the AU detection task, where, again, we show the models' performance on the upper face AUs $(1, 2, 4, 6, 7)$. To form the positive and negative class for each AU, we adopted the same approach as for the MMI dataset.

As input features from the CK+ and AFEW datasets, we used the locations of 49 facial points, and for the MMI dataset we used the locations of 20 facial points, all provided by the database creators. For the GEMEP-FERA, we used the locations of 49 facial points extracted from target images sequences using the appearance-based facial tracker in [40]. Fig.2 depicts the used facial points from each dataset. The pre-processing of the input features (i.e., the facial points) was performed by first applying Procrustes analysis to align the facial points to the mean faces of the datasets. This is important in order to reduce the effects of head-pose and subject-specific variation. We then applied PCA to reduce the feature size, retaining 97% of energy, resulting in 26, 18, 20 and 20 dimensional feature vectors for the CK+, AFEW, MMI and GEMEP-FERA datasets, respectively. The obtained features were used as inputs to tested models. We set in all our experiments the number of hidden states $C = 3$ for both ordinal and nominal classes. This number of states corresponds to the temporal phases of expression development (neutral-onset/offset-apex). As evaluation measures, we report the F-1 score and the average classification accuracy.

The learning of the VSL-CRF model parameters was performed by randomly initializing the parameters of the nominal/ordinal score functions. In order to avoid the subgradients falling into a local minimum (by diverging to either nominal or ordinal models), we performed random assignments of the training sequences to either nominal or ordinal classes during the first 10 iterations of the LBFGS optimization (see Sec.III-C). After this, we applied the max-pooling of the models with the proposed subgradient approach until the convergence of the objective function. The regularization parameters $\lambda_{n(o)}$ were set using a grid-search validation procedure on the training set. For the competing models, H-CRF and H-CORF, we used our own implementation. The initial model parameters were set using the same approach as in the VSL-CRF. As the baseline for the sequence classification we also include the results obtained by first applying SVMs (with the RBF kernel), trained/evaluated per frame, followed by the majority voting. We refer to this approach as SVM-SB. To compare the performance of target models with the state-of-the-art models for each of target tasks (emotion recognition and AU detection), we report the results from the original papers, as detailed below. Although these are not directly comparable, as they used different settings (per-frame/sequence and number of folds, as well as different features), we show these results for completeness. However, our main goal is to demonstrate the benefits of using the variable latent states (VSL-CRF) compared to when only nominal (H-CRF) or ordinal (H-CORF) states are used.



(a) CK+ [34] (b) AFEW [35] (c) MMI [36] (d) FERA [37]

Fig. 2: Sample images with the used facial points from four different datasets.

Method (7 emotions)	F-1 [%]	Accuracy [%]
EmotiW [41]	—	27.3
STM-ExpLet [7]	—	31.7
SVM-SB	26.3	31.2
H-CRF	19.7	22.6
H-CORF	22.4	27.4
VSL-CRF	**28.1**	**32.2**

TABLE II: AFEW dataset.

These three models were all evaluated using exactly the same setting.

Table I shows the results for facial expression recognition from the CK+ dataset. Note that some of the methods compared use different number of folds. Specifically, Bartlet et al. [38] applied the SVM classifier to each image frame, and the label of a sequence is decided by a majority vote using a 4 fold cross validation. ITBN [39] performed a 15-fold cross validation. This method is based on the Interval Temporal Bayesian Network, which models spatial and temporal relations of facial expressions. TMS [12] applied a 4-fold cross validation. This approach uses Latent-Dynamic CRFs [22], where the emotion recognition is performed per-frame. PO-HCRF9 (partially observed H-CRF) [13] used a 5-fold cross-validation. Note that in this method, some states are observed during training and represent activations of AUs. Lastly, we compare our method to the state-of-the-art method for target task, STM-ExpLet [7], where the same experimental setting is used as in our experiments. We see from Table I that the proposed method outperforms the other methods. Similar observations can be made from Table II, which shows the results for the AFEW dataset. EmotiW [41] is the baseline, obtained using SVMs, from the dataset paper. Note that here all the methods achieve low recognition results, as the target dataset is very challenging, and, also, the emotion labels were obtained in a semi-automatic way. Still, the proposed VSL-CRF largely outperforms H-CRF/H-CORF. Interestingly, the baseline SVM-SB approach achieves similar results to our method. This is mainly because of the lack of temporal structure in the used data (some parts of the videos contained the target-expression-unrelated content), as well as because of the uncertainty in the emotion labels.

Table III shows results for AU detection from the GEMEP-FERA and MMI datasets. The STM [4] is a transductive learning method, which personalizes the SVM classifier by re-learning its parameters for each test person. Despite this adaptation, it still fails to reach the full performance of the VSL-CRF model. This is attributed to the fact that the STM does not model the temporal dynamics. However, H-CRF/H-CORF models do not outperform STM on all AUs. Hence, the assignment of both types of latent states, as done in the VSL-CRF model is critical for achieving superior performance on this task. This is also reflected in results obtained on the MMI dataset. Specifically, the sequence based methods (SVM-SB, H-CRF, H-CORF and VLS-CRF) models largely outperform the existing frame-based methods (PFFL [26] and FFD [24] on the AU detection task. Yet, the proposed VSL-CRF outperforms these models on both datasets. LPQ-TOP [42] is also a frame-based model that focuses on novel feature definition for AU detection. Although

	GEMEP-FERA					MMI						
AU	STM [4]	SVM-SB	H-CRF	H-CORF	VSL-CRF	PFFL [26]	FFD [24]	LPQ-TOP [42]	SVM-SB	H-CRF	H-CORF	VSL-CRF
1	68.1 (—)	69.4 (63.8)	70.0 (70.1)	70.0 (66.6)	**72.9 (73.1)**	70.0 (—)	72.7 (—)	85.6 (—)	81.8 (85.9)	87.4 (87.5)	87.5 (87.5)	**91.7 (91.2)**
2	65.5 (—)	69.7 (65.1)	70.5 (74.3)	73.5 (74.3)	**85.9 (85.6)**	62.3 (—)	72.7 (—)	79.4 (—)	79.6 (79.8)	87.4 (87.5)	78.5 (78.4)	**87.5 (87.8)**
4	43.3 (—)	58.8 (56.7)	61.2 (61.4)	61.3 (61.4)	**70.0 (70.6)**	64.0 (—)	67.3 (—)	**81.2** (—)	58.8 (61.6)	57.0 (57.1)	63.8 (64.4)	69.5 (70.7)
6	71.6 (—)	76.9 (63.1)	63.7 (63.7)	70.1 (70.8)	**79.1 (79.1)**	63.4 (—)	73.7 (—)	**87.2** (—)	63.9 (68.2)	67.0 (67.3)	63.3 (63.5)	78.7 (78.8)
7	66.2 (—)	60.3 (63.1)	55.8 (55.7)	58.1 (58.1)	**74.3 (74.4)**	39.2 (—)	36.4 (—)	80.9 (—)	77.9 (84.0)	66.1 (66.6)	86.7 (86.6)	**87.7 (87.6)**
AVG	62.9 (—)	67.0 (66.1)	64.2 (65.0)	66.6 (66.2)	**76.4 (76.6)**	59.8 (—)	64.6 (—)	82.6 (—)	72.4 (75.9)	73.0 (73.2)	76.0 (76.1)	**83.0 (83.2)**

TABLE III: AU detection from the GEMEP-FERA and MMI datasets. The numbers shown represent the F-1 (Accuracy) scores in % for each method.

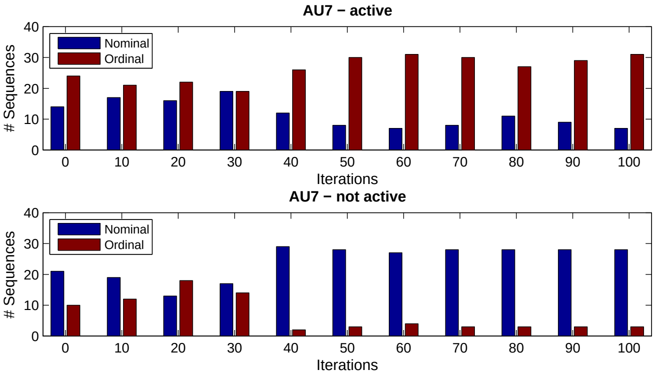


Fig. 3: AU7 (lid tightener) from the GEMEP-FERA dataset. The figures show the number of training sequences from the AU7 active class (upper) and AU7 not active class (lower) being assigned nominal/ordinal latents states during the model learning.

the average results are slightly worse than those achieved by the proposed VSL-CRF, they are not directly comparable. While the work in [42] is based on a static per-frame classification, the proposed work focuses on modeling spatio-temporal dynamics of sequence data. The high F-1 score achieved by both methods demonstrates the importance of both tasks, modeling the dynamics and the features of facial expressions. Nevertheless, we applied the same validation setting for the sequence-based methods. Hence, the results by these methods are directly comparable and show that the proposed VSL-CRF outperforms the other methods on all used datasets. This is mainly because of its ability to select the optimal latent states (nominal or ordinal) for representing the target sequences.

We also inspect how the assignment of the nominal/ordinal states is attained during learning and inference in the proposed VSL-CRF model. For this, we show the model's behavior in the task of detection of AU7 (lid tightener) from the GEMEP-FERA dataset. Fig.3 depicts the model's selection of ordinal/nominal states for the target sequences being labeled as AU7 active or not active. At the beginning (iteration 0), the model is initialized by random assignment of sequences to either nominal or ordinal states (see Sec.III-C). As can be seen, for this fold, the model converges by assigning to the majority sequences of the positive class (AU7 active) the ordinal states, and the nominal states to the negative class (AU7 not active). It is important to mention that during its learning, the model does not disregard neither nominal nor ordinal node potentials for each class, but rather selects the optimal parameters for the score functions of both.

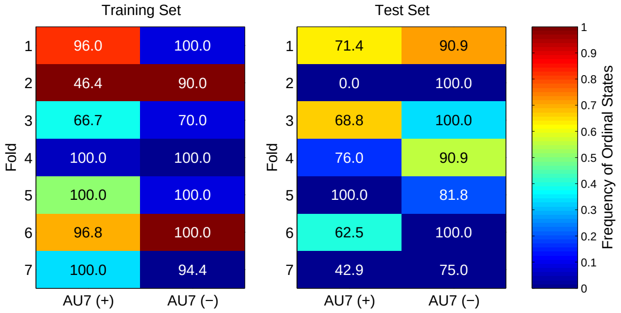


Fig. 4: The proportion of the training/test sequences being assigned ordinal (nominal) states (depicted by different color) in the learned VSL-CRF model for detection of AU7 from the GEMEP-FERA dataset. The numbers shown represent the true positives (TP) and true negatives (TN) in %, and are computed per fold.

This is further illustrated in Fig.4, showing the model's selection of the type of the latent states per fold (each training fold containing sequences of 6 subjects, and test fold containing sequences of 1 test subject). From the colors depicting the proportion of the nominal/ordinal sequences, we conclude that this choice depends largely on the training/test subjects in each fold. For instance, in the case of fold 1, the learning of which is also depicted in Fig.3, although the negative class is mainly nominal on the training set, this is not the case on the test set. For the negative test class, the ordinal latent states are predominant, with the true negatives (the depicted numbers in Fig.4) being relatively high (90.9%). Thus, the model employs ordinal states for this class as they turn out to fit better the negative class of this test subject. On the other hand, for fold 2, both classes are mainly ordinal on the training set, while both exhibiting nominal type on the test set. From the true positives for both the training and test sets, we observe that in the case of this fold, the model failed to fit the positive class well, and consequently, poorly performed inference of the test data. We attribute this to the model falling into local minimum during parameter optimization, which adversely affected its generalization to the positive class of the test set. Note also that for fold 2, the negative test class is mainly nominal, despite the fact that the corresponding training class is ordinal. A possible reason for this is the large difference in facial features between the training and test subjects, as well as disbalance in examples of the positive/negative class. The remaining folds can be analyzed in a similar manner. We plan to investigate such behavior of the VSL-CRF model in more detail in our future work.

V. Conclusions

In this paper, we proposed a novel Latent Conditional Random Field model for dynamic facial expression recognition and AU detection. By allowing the structure of the latent states of target classes to vary for each target sequence, we obtained the model that can better discriminate between different facial expressions than the existing models that restrict their latent states to have the same and pre-defined structure for all classes (nominal or ordinal). We showed on four facial expression datasets that the proposed model outperforms the state-of-the-art sequence- and frame-based methods for facial expression recognition and AU detection.

Acknowledgments

This work has been funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). The work of Ognjen Rudovic is also funded in part by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA). The work of Vladimir Pavlovic has been funded by the National Science Foundation under Grant no. IIS0916812.

References

[1] M. Pantic and L .J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *TPAMI*, 22(12):1424–1445, 2000.

[2] J. F. Cohn and P. Ekman. Measuring facial actions. In *The New Handbook of Methods in Nonverbal Behavior Research, Harrigan, J.A., Rosenthal, R. & Scherer, K., Eds.*, pages 9–64. 2005.

[3] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, 2002.

[4] W. Chu, F. Torre, and J. F Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR*, pages 3515–3522. IEEE, 2013.

[5] S. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and Trevor Darrell. Hidden conditional random fields for gesture recognition. *CVPR*, pages 1097–1104, 2006.

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

[7] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, pages 1749–1756, 2013.

[8] O. Rudovic, V. Pavlovic, and M. Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. *IEEE CVPR*, pages 2634–2641, 2012.

[9] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *TPAMI*, 2014.

[10] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2):210–220, 2006.

[11] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *transactions on pattern analysis and machine intelligence*, 29(10):1848, 2007.

[12] S. Jain, Changbo Hu, and J.K. Aggarwal. Facial expression recognition with temporal modeling of shapes. In *ICCV'W*, pages 1642–1649, 2011.

[13] K. Chang, T. Liu, and S. Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. In *CVPR*, pages 533–540. IEEE, 2009.

[14] M. Held, P. Wolfe, and H. P Crowder. Validation of subgradient optimization. *Mathematical programming*, 6(1):62–88, 1974.

[15] Z. Zeng, M. Pantic, G.I Roisman, and T.S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *TPAMI*, 31(1):39–58, 2009.

[16] C. Shan, S. Gong, and P. W McOwan. Conditional mutual infomation based boosting for facial expression recognition. In *BMVC*, 2005.

[17] M. S. Bartlett, G . Littlewort, I. Fasel, and J. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *CVPR*, 2003.

[18] M. Pantic and L. JM Rothkrantz. Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(3):1449–1461, 2004.

[19] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[20] J. D. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML '01, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.

[21] L. Shang and K. Chan. Nonparametric discriminant hmm and application to facial expression recognition. In *CVPR*, pages 2090–2096, 2009.

[22] N. Sebe, M.S Lew, Y. Sun, I. Cohen, T. Gevers, and T.S Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.

[23] M. F. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *CVPRW*, 2005.

[24] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *TPAMI*, (11):1940–1954, 2010.

[25] B. Jiang, M. Valstar, and M Pantic. Facial action detection using block-based pyramid appearance descriptors. In *Social Compug*, 2012.

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

[27] L. van Maaten and E. Hendriks. Action unit classification using active appearance models and conditional random fields. *Cognitive Processing*, 13(2):507–518, 2012.

[28] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *Transactions on Cybernetics*, 44(2):161–174, 2014.

[29] O. Rudovic, V. Pavlovic, and M. Pantic. Kernel conditional ordinal random fields for temporal segmentation of facial action units. *IEEE ECCV'W*, 2012.

[30] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, pages 282–289, 2001.

[31] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Stat. Society. Series B*, 42:109–142, 1980.

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

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

[34] P. Lucey, Jeffrey F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion expression. In *CVPRW*, 2010.

[35] Abhinav Dhall et al. Collecting large, richly annotated facial-expression databases from movies. *MultiMedia, IEEE*, pages 34–41, 2012.

[36] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. Intl Conf. Language Resources and Evaluation*, pages 65–70, 2010.

[37] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *FG 2011*, pages 921–926, 2011.

[38] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *CVPRW'*, 2005.

[39] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *CVPR*, pages 3422–3429. IEEE, 2013.

[40] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPRW*, pages 896–903, 2013.

[41] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *ACM*, pages 509–516, 2013.

[42] B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011.