Image and Vision Computing 58 (2017) 25–37

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

# Variable-state Latent Conditional Random Field models for facial expression analysis☆

Robert Walecki[a,*], Ognjen Rudovic[a], Vladimir Pavlovic[b], Maja Pantic[a]

[a]Computing Department, Imperial College, London, UK
[b]Department of Computer Science, Rutgers University, USA

ARTICLE INFO

*Article history:*
Received 20 November 2015
Accepted 23 April 2016
Available online 22 June 2016

*Keywords:*
Facial expression
Action unit
Conditional Random Fields
Sequence classification
Segmentation

ABSTRACT

Automated recognition of facial expressions of emotions, and detection of facial action units (AUs) from videos depends critically on modeling of their dynamics. Some of these dynamics are characterized by changes in temporal phases (onset-apex-offset) and intensity of emotion expressions and AUs. The appearance of these changes may vary considerably among subjects, making the recognition/detection task very challenging. The state-of-the-art Latent Conditional Random Fields (L-CRF) framework allows us to efficiently encode these dynamics through the latent states accounting for the temporal consistency in emotion expression and ordinal relationships between its intensity levels. These latent states are typically assumed to be either unordered (nominal) or fully ordered (ordinal). Yet, while the video segments containing activation of the target AU may better be described using ordinal latent states (corresponding to the AU intensity levels), the segments where this AU does not occur, may better be described using unordered (nominal) latent states. To address this, we propose the variable-state L-CRF (VSL-CRF) model that automatically selects the optimal latent states for the target image sequence, based on the input data and underlying dynamics of the sequence. To reduce the model overfitting, we propose a novel graph-Laplacian regularization of the latent states. We evaluate the VSL-CRF on the tasks of facial expression recognition using the CK+ dataset, and AU detection using the GEMEP-FERA and DISFA datasets, and show that the proposed model achieves better generalization performance compared to traditional L-CRFs and other related state-of-the-art models.



## 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. Machine understanding of facial expressions could revolutionize user interfaces for artifacts such as robots, mobile devices, cars, and conversational agents [1]. Other valuable applications are in the domain of medicine and psychology, where it can be used to improve medical assistance as well as develop automated tools for behavioral research [2]. Therefore, automated analysis of facial expressions has attracted a significant research attention [3, 4]. Facial expressions (FE) 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 [5]. The message judgment approach aims to directly decode the meaning conveyed by a facial display (e.g., in terms of the six basic emotions). Instead, the sign judgment approach aims to study the physical signal used to transmit the message (such as raised cheeks or depressed lips). To this end, the *Facial Action Coding System* (FACS) [6] is used as a gold standard. It is the most comprehensive, anatomically-based system for encoding facial expressions by describing the facial activity based on the activations of 33 AUs. These AUs, individually or in combinations, can describe nearly all-possible facial movements [6, 7].

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 [3]. However, recognizing facial expressions from videos (i.e., image sequences) is more natural and has proved to be more effective [1, 8]. This is due to the fact that facial expressions can better be described as a dynamic process that evolves over time. For instance, facial expressions of emotions and AUs undergo a transition of their temporal phases (onset-apex-offset) during the expression development. Similarly, the activation of AUs spans different time intervals that reflect variation in their intensity, as described by FACS. Several works in the field (e.g., [1–3]) have emphasized the importance of modeling these dynamics for increasing the recognition performance in the target tasks compared to the static methods (see also [8]).

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author at: Queens Gate 180, SW7 2AZ, U.K.
*E-mail address:* r.walecki14@imperial.ac.uk (R. Walecki).

http://dx.doi.org/10.1016/j.imavis.2016.04.009
0262-8856/

Most of the state-of-the-art approaches for modeling facial expression dynamics are based on variants of Dynamic Bayesian Networks (DBN) (e.g., Hidden Markov Models (HMM) [9]) and on Conditional Random Fields (CRF) [10]). These methods are detailed in Section 2.1. In what follows we focus on hierarchical extensions of CRF [2, 11, 12, 13], as they are directly related to the model proposed in this paper. These methods can be cast as variants of the CRF called Latent CRF (L-CRF) [14], and they have also been successfully used for other computer vision problems (e.g. gesture recognition [14] and human motion estimation [15]). In the context of facial expressions, L-CRF 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., an emotion category). A typical representative of these models is the Hidden CRF (H-CRF) [14, 16, 17, 18], used for facial expression recognition of six basic emotions. Apart from temporal constraints imposed on its latent states, this model fails to account for the ordinal relationships between the latent states. However, this may be important if the aim is to encode intensity of target events as it is the case with encoding the intensity of facial expressions. To this end, the recently proposed Hidden Conditional Ordinal Random Field (H-CORF) model [11, 12] imposes additional constraints on the latent states of modelled events by exploiting their ordinal relationships. Specifically, this model implicitly enforces the latent states (e.g. emotions) to correlate with their temporal phases (or intensity) by representing them on an ordinal scale. This, in turn, results in the model with fewer parameters, which is less prone to overfitting, and, thus, able to discriminate better between events (e.g. facial expressions of different emotions [11, 12]).

However, in the L-CRF 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 unconstrained nominal model provides a better fit to the data. For example, in recognition of emotion-specific expressions, we expect the latent states used to model the activation of facial expressions of target emotion class (e.g., happiness) to be correlated with its temporal phases defined on an ordinal scale (neutral $<$ onset $<$ apex). Similarly, for an AU activation, the latent states should be correlated with its intensity levels, as defined on the Likert scale using FACS (i.e., neutral $<$ A $<$ B $<$ C $<$ D $<$ E). On the other hand, image sequences of the negative class, i.e., containing a neutral face (without facial activity) or a mix of other non-target facial expressions (different emotions or AUs), are expected to model best using nominal states. This is due to the lack of the ordinal structure as well as high variability (activations of various non-target AUs) in such data. We can even go a step further by assuming that the nature of the latent states depends not only on the type of the emotion/AU class (active vs inactive), but that it can also vary for each image sequence of the target classes. For instance, in case of noisy image features (due to the tracking errors in the case of facial landmarks) and due to differences in facial expressiveness of different subjects, resulting in subject-specific features.

In these cases, the ordinal relationships could be altered and, thus, modeling of the ordinal latent states may not be flexible enough to account for the increased levels of variation in the data. To mitigate this, the model should automatically infer what type of the latent states should be used for modeling the dynamics of the input/output data. To this end, we generalize the L-CRF models by relaxing their assumption that the latent states within the target sequence need only be nominal or ordinal. Specifically, we introduce a novel latent variable within the L-CRF framework, the state of which defines the type of latent states that are best suited for target image sequences. The learning in the proposed model is performed using two newly defined approaches based on max-polling of the latent states and the Expectation–Maximization (EM) algorithm. To reduce potential redundancy in the modeling of the underlying dynamics of facial expressions, we propose the graph-Laplacian regularization of the model parameters that is defined directly on posterior distributions of the latent states.

The contributions of the proposed work can be summarized as follows:

1) We introduce a novel Variable-state L-CRF (VSL-CRF) model for classification of image sequences that, in contrast to existing L-CRF models, has flexibility to use either nominal or ordinal latent states for modeling the underlying dynamics of target events. Also, the proposed model selects automatically the optimal latent states for each target sequence.
2) We propose two novel learning algorithms based on max-pooling and the EM-like learning of the latent states, as well as graph-Laplacian regularization of the model parameters, for efficient training of the proposed VSL-CRF model. This results in a model that is less prone to overfitting than those based on maximum-likelihood learning (ML) approach as in L-CRF models (H-CRF and H-CORF).
3) We show on three publicly available datasets (CK+, GEMEP-FERA and DISFA) that the VSL-CRF model achieves superior performance in classification of facial expressions. This is due to its ability to learn the well underlying dynamics of the target facial expression.

The rest of the paper is organized as follows. Section 2 describes the recent advances in the sequence- and frame-based classification of facial expressions of emotions and AU detection. Section 3 introduces the proposed methodology. Section 4 describes the conducted experiments and presents the evaluation results, and Section 5 concludes the paper.

## 2. Related work

### 2.1. Facial expression recognition

Facial expression recognition methods can be categorized into the static and dynamic approaches (see [8] for a detailed overview). The static approach attempts the expression recognition from a single image (typically, the apex of the expression) [19–21]. For example, Zeng et al. [22] proposed a two-stage multi-task sparse learning framework to efficiently locate the most discriminative facial patches for the expression classification. The SVM classifier is then used to classify the patches into the six basic emotion categories. The approach in [23] exploits ensemble of features comprising of Hierarchical Gaussianization (HG), Scale Invariant Feature Transform (SIFT) and Optic Flow, followed by the SVM-based classification of emotion expressions.

However, a natural facial event such as facial expression of an emotion is dynamic, i.e., it evolves over time by (typically) starting from a neutral expression, followed by its onset, apex, and then the offset, followed by the neutral expression again. For this reason, facial expression recognition from videos is more common than from static images. Although some of the static methods use the features extracted from a window around the target frame, in order to encode dynamics of facial expressions, models for dynamic classification provide a more principled way of doing so. As we mentioned in Section 1, most of the dynamic approaches to classification of facial expressions are based on variants of DBNs such as HMMs and CRFs. For example, Shang et al. [24] trained independent HMMs for each emotion category, and then performed emotion classification by comparing the likelihoods of the emotion-specific HMMs. However, discriminative models based on CRFs [17, 18, 25] have been shown to be more effective for the facial

expression classification. Furthermore, Wang et al. [26] have shown that capturing more complex time-dependences in the data (beyond the first order dependences as done in linear-chain CRFs) can enhance the facial expression classification performance. Similarly, Jain et al. and Sebe et al. [17, 25] used a generalization of the linear-chain CRF model, a Hidden Conditional Random Field (H-CRF) [14], with additional layer of (hidden) variables 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 facial expressions of different emotion categories (based on comparisons with SVMs). Another modification of H-CRF, named partially-observed H-CRF, was proposed in [18], 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-occurrences. In contrast to these models, which still perform per-frame classification of target expressions, [11, 12] proposed the Hidden Conditional Ordinal Random Field (H-CORF) models for the sequence-based classification of facial expressions of emotions and their temporal phases (onset-apex-offset) simultaneously. These models encode ordinal relationships between the temporal phases of emotion expression using either supervised or unsupervised learning of the latent states (corresponding to the temporal phases). The authors showed that improved facial expression recognition can be achieved due to the ordinal modeling of the latent states, with the supervised modeling of the latent states (i.e., using the labels for the temporal phases of emotion expression) outperforming the unsupervised modeling, as expected in this task. Nevertheless, the main limitation of the models listed here is that they restrict their latent states to be either nominal (H-CRF) or ordinal (H-CORF), which may be suboptimal in some cases, as discussed in Section 1.

### 2.2. Facial AU detection

As for facial expression recognition, two main approaches are typically adopted for AU detection: static and dynamic modeling. In the former, image features are extracted from each frame and then fed into a static classifiers such as SVM or AdaBoost [27] specifically designed for detection of each AU independently. A more advanced static AU detector, named The Selective Transfer Machine (STM) [28], has shown great improvements over standard SVMs in the target task. It 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. However, a limitation of this approach is that the re-learning of the target AU detectors has to be performed for each test subject. The modified correlation filter (MCF) [29] is also an approach similar in spirit to SVMs and correlation filters, but with the key difference of optimizing only a single hyperplane. This results in more robust AU detection compared to standard SVMs when sequence-level AU labels are used for the frame-based AU detection. The authors of [30] proposed a multi-kernel-learning (MKL) approach to AU detection, where they investigate the fusion of different appearance-based image features via the sum of histogram-based kernel functions. These kernels are then used in the SVMs trained for each AU. To include the temporal information, the authors extract features within AU-specific windows around the image frames used for detection of target AUs. Zhu et al. [31] proposed a multi-task feature learning (MTFL) method for joint AU detection. The MTFL approach and Bayesian networks are used to model AU dependences at both feature and label level, and, thus, perform joint AU detection in a probabilistic fashion. Likewise, Zhang et al. [32] introduces the lp-norm regularization to the MKL, in order to fuse multiple features (using various kernels) and account for the AU-dependencies. Bayesian graphical models were also used to encode sparsity and statistical co-occurrence of AUs [33] for their joint modeling.

While the methods listed above focus on finding the most discriminative feature representations and/or on inference methods for joint AU detection, they fail to account for temporal information, i.e., AU dynamics. Methods that do so attempt using either temporal image features [34, 35] or DBN-based models such as HMMs [7] and CRFs [36]. In general, these works perform either majority voting using the static detection [27], 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) [7, 37]. Other temporal models are based on Ordinal CRFs that have been proposed for modeling of AU temporal phases [38], and their intensity [2], however, they do not perform AU detection. Another approach, termed Cascade of Tasks (CoT) [39], is trained on sequences and applies segment-based detection of AUs. This approach is a combination of three algorithms for static-frame-level-detection, segment-level-detection and transition-level detection. The Interval Temporal Bayesian Networks [26] (ITBN) have also been proposed to capture complex temporal relations among facial events, and for AU detection. The network also represents the spatial dependences among the facial events with a larger variety of time-constrained relations.

Note that the above-mentioned approaches for facial expression recognition and AU detection 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 under-constrained (e.g., H-CRF[14]) or over-constrained (H-CORF[11]), which limits their representational power. In relation to the state-of-the-art methods, the proposed VSL-CRF model focuses on two key aspects of the facial expression recogniton/ AU detection: (i) modeling of their temporal dynamics (via novel latent states of the L-CRF models) to improve the recognition/detection performance of existing graph-based dynamic models for the target task. (ii) The application of the model to the sequence-based classification and frame-based detection of facial expressions of emotions and AUs. In the following, we introduce the proposed methodology.

## 3. Methodology

In this section, we first give a short introduction to ordinal and nominal CRFs, and their L-CRF extensions. We then introduce the VSL-CRF method that generalizes these approaches. Lastly, we introduce different methods for the model optimization, including the posterior regularization of the latent states.

### 3.1. Notation

We consider a $K$-class classification problem, where we let $y \in \{1, \dots, K\}$ 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 $x = \{x_1 \dots x_T\} \in T \times D$, serves as input covariates for predicting $y$ and $h = (h_1, \dots, h_T)$. 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 $\mathcal{D} = \{(y^i, x^i)\}_{i=1}^N$, which are i.i.d. samples from an underlying but unknown distribution.

### 3.2. Conditional Random Fields (CRF)

CRFs [40] are a class of log-linear models that represent the conditional distribution $P(h|x)$ as the Gibbs form clamped on the observation $x$:

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

Here, $Z(x;\boldsymbol{\theta}) = \sum_{h\in\mathcal{H}} e^{s(x,h;\boldsymbol{\theta})}$ 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) $s(x,h;\boldsymbol{\theta})$. Note that in this model, the states $h$ are observed and they represent the frame labels.

#### 3.2.1. Linear chain Conditional Random Fields (CRF)

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 $\boldsymbol{\Psi}_r^{(V)}(x,h_r)$ and the edge features by $\boldsymbol{\Psi}_e^{(E)}(x,h_r,h_s)$. By letting $\boldsymbol{\theta} = \{v,u\}$ be the parameters of the node and edge potentials, respectively, $s(x,h;\boldsymbol{\theta})$ can then be written as the sum:

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

Although the representation in Eq. (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 $v^\top \boldsymbol{\Psi}_r^{(V)}(x,h_r) \to \boldsymbol{\Gamma}_r^{(V)}(x,h_r)$, where

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

The **nominal** node potential is then obtained by using the multinomial logistic regression (MLR) model [14]:

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

where $f_n(x,c) = \beta_c^T\cdot[1,x]$, for $c = 1,\ldots,C$, and $\beta_c$ is the separating hyperplane for the $c$-th *nominal* state of the target class. By plugging the likelihood function in Eq. (4) into the node potential in Eq. (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 [41], and has the form:

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

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 $b_0 = -\infty \le \cdots \le b_C = \infty$, with each bin corresponding to one of the *ordinal* states $c = 1,\ldots,C$ in the model. The ranking likelihood in Eq. (5) is constructed by contaminating the ideal model (see [42] for details) with Gaussian noise with standard deviation $\sigma$.

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

Again, by plugging the likelihood function in Eq. (5) into the node potential in Eq. (3), we obtain the node features of the Ordinal CRF (CORF) model [42].

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

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

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.

#### 3.2.2. Latent Conditional Random Fields (L-CRFs)

While the CRFs introduced in the previous section aim at modeling/decoding of the state-sequence within a single class, the framework of L-CRFs [14, 43] 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 $h$ are now treated as unknown. Formally, L-CRFs combine the score functions of $K$ CRFs, one for each class $y = \{1,\ldots,K\}$, within the following score function:

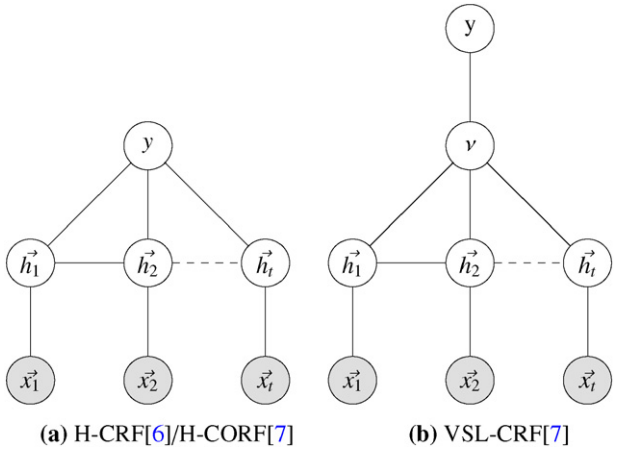
$$s(y,x,h;\boldsymbol{\Omega}) = \sum_{k=1}^{K} I(k = y)\cdot s(x,h;\theta_y), \tag{7}$$

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

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

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

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




**(a)** H-CRF[6]/H-CORF[7] **(b)** VSL-CRF[7]

**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 $\mathrm{x} = \{x_1,\ldots,x_T\}$ 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.

Evaluation of the class-conditional $P(y|x)$ depends on the partition function $Z(x) = \sum_k Z_k(x) = \sum_k \sum_h \exp(s(k,x,h))$, and the class-latent joint posteriors $P(k,h_r,h_s|x) = P(h_r,h_s|x,k)\cdot P(k|x)$. Both can be computed from independent consideration of $K$ individual CRFs. The model with the *nominal* node potentials in the score function in Eq. (9) is termed Hidden CRF (H-CRF) [14]. Likewise, the model with the *ordinal* node potentials is termed Hidden CORF (H-CORF) [11].

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 Eq. (9). Furthermore, to avoid the constrained optimization in H-CORF (due to the order constraints in parameters **b** of the ordinal node potentials), the displacement variables $\gamma_c$, where $b_j = b_1 + \sum_{k=1}^{j-1} \gamma_k^2$ for $j = 2,\ldots,C-1$ are introduced. So, **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 log-partition function (log-sum-exp of nonlinear concave functions), their log-likelihood objective is bounded below by 0 and are both smooth functions. For this, the standard quasi-Newton (such as Limited-memory BFGS) gradient descent algorithms are typically used to estimate the model parameters (we use the former). The model parameters for H-CRF are given by $\theta_y^{(n)} = \beta_1,\ldots,\beta_C$, where $C$ is the number of nominal latent states for class $y = \{1,\ldots,K\}$. Likewise, for H-CORF we have $\theta_y^{(o)} = \{b_1,\gamma_1,\ldots,\gamma_{C-2},\sigma\}$ for each class in $y$.

### 3.3. 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 (latent states) within each sequence. 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.

#### 3.3.1. VSL-CRF: model

**Definition 1** (Variable-state Latent CRF)**.** Let $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_K)$ be a vector of symbolic states or labels encoding the nature of the latent states $h^\nu$ of the $i$-th sequence, $i = 1,\ldots,N_y$ from class $y = (1,\ldots,K)$, 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:

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

where the nominal ($s(x,h;\theta_y^n)$) and ordinal ($s(x,h;\theta_y^o)$) score functions represent the sum of the node and edge potentials, as given by Eqs. (3) and (6), respectively. Then, the full conditional probability of the VSL-CRF model is given by:

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

$$Z(x) = \sum_{k,h,\nu}\exp(s(k,x,h,\nu)) \tag{12}$$

Note that, in contrast to L-CRF models introduced in Section 3.2, the VSL-CRF performs also integration over the latent variable $\nu$, the state of which (ordinal or nominal) defines the type of the latent states for each sequence of facial expressions. The definition of the VSL-CRF in Eq. (11) allows it to simultaneously fit both ordinal and nominal latent states to each sequence, which may result in the model overfitting. In the following, we introduce two novel learning strategies in order to avoid over-parametrization of the model, i.e., to prevent the model from using redundant nominal and/or ordinal latent states during inference of target sequences.

#### 3.3.2. VSL-CRF: learning and inference

*3.3.2.1. Max-pooling of latent states.* The first learning strategy that we propose constraints the latent states to take either nominal or ordinal sequence of latent states per target sequence. This is different from H-CRF/H-CORF where the latent states can be either nominal/ordinal for each and every target class and the sequence. Formally, the conditional probability in Eq. (11) is now given by:

$$P(y|x) = \frac{\max_\nu\left(\sum_h \exp(s(y,x,h,\nu))\right)}{Z(x)} \tag{13}$$

$$Z(x) = \sum_k \max_\nu\left(\sum_h \exp(s(k,x,h,\nu))\right) \tag{14}$$

The key aspect of this approach is that now the type of the latent states is explicitly constrained to either nominal or ordinal. This, in turn, leads to the following (regularized) loss function of the VSL-CRF model (further in the text, we denote this model as **VSLm**):

$$RLL(\boldsymbol{\Omega}) = -\sum_{i=1}^{N}\log P(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\Omega}) + \lambda_{n(o)}||\theta_{k=1..K}^{n(o)}||^2, \tag{15}$$

where $\boldsymbol{\Omega} = \{\theta_k^n,\theta_k^o\}_{k=1}^K$. We introduce $L$-2 regularization over the parameters of the nominal/ordinal score functions, the effect of which is controlled by $\lambda_n/\lambda_o$, which are found using a validation procedure.

Unfortunately, the objective function of the VSLm model is both non-convex and non-smooth because of the *max* function in its conditional distribution. Therefore, the gradients of the objective in Eq. (15) w.r.t. the parameters $\boldsymbol{\Omega}$ cannot be directly computed. Yet, the nominal/ordinal score functions are both sub-differentiable. We use this property to construct the sub-gradient [44] of the VSLm objective at $\boldsymbol{\Omega}$. Essentially, this boils down to computing the following sub-gradients

$$\partial\boldsymbol{\Omega} = \nabla\max_\nu\left(\sum_h \exp(s(k,\mathbf{x},\mathbf{h},\nu))\right),\ k = 1,\ldots,K,$$

which are further given by

$$\begin{cases} \partial\boldsymbol{\Omega} = \nabla\sum_h \exp(s(\mathbf{x},\mathbf{h},\theta_k^n)), \\ \quad \textit{if } \sum_h \exp(s(\mathbf{x},\mathbf{h},\theta_k^n)) > \sum_h \exp(s(\mathbf{x},\mathbf{h},\theta_k^o)) \\ \partial\boldsymbol{\Omega} = \nabla\sum_h \exp(s(\mathbf{x},\mathbf{h},\theta_k^0)), \quad \textit{otherwise}. \end{cases}$$

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, $\max_\nu\left(\sum_h \exp(s(k,\mathbf{x},\mathbf{h},\nu))\right)$ is differentiable and has the gradient $\partial\theta_k^n = \nabla\sum_h \exp(s(\mathbf{x},\mathbf{h};\theta_k^n))$, 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 [11] for more details).

*3.3.2.2. Fully integrated out latent states.* The benefits of the VSLm approach are that it prevents the VSL-CRF model from redundant parametrization of the VSL-CRF model that can easily lead to the model overfitting. However, the sub-gradient optimization approach can easily get trapped in local minimum when searching for the model parameters due to the gradient 'switching' caused by the *max* function in the objective. To this end, we also employ a learning strategy where both types of the latent states (ordinal and nominal) are fully integrated out, which can be solved using the standard gradient descent optimization as in the existing L-CRF models. Of course, the downside is that we may end up with over-parametrization of the target sequences. To remedy this, in addition to direct optimization of the conditional probability, we also introduce an EM approach to the parameter learning.

$$P(y|x) = \frac{\sum_{\nu}\left(\sum_{h}\exp(s(y,x,h,\nu))\right)}{Z(x)} \tag{16}$$

In the proposed EM learning strategy, we exploit the hierarchy in the VSL-CRF model, which allows us to integrate out the latent states $h$ and the indicator variable $\nu$ in an alternating fashion. Note that no empirical studies that investigate the performance of EM vs. the direct optimization in the context of L-CRFs have been reported so far. Furthermore, we introduce novel posterior regularization (see Section 3.4) in the objective function of these approaches, with the aim of implicitly enforcing the model to select either nominal or ordinal latent states for each target sequence during learning[2] . Formally, the objective function is given by:

$$RLL(\mathbf{\Omega}) = -\sum_{i=1}^{N}\log P(\mathbf{y}_i|\mathbf{x}_i;\mathbf{\Omega}) + \lambda_{n(o)}||\theta^{n(o)}_{k=1..K}||^2 + \lambda_p\sum_{\nu'}\mathcal{R}_{\nu'} \tag{17}$$

where $P(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\Omega})$ is defined by Eq. (16), and $\lambda_p$ controls the strength of the posterior regularization defined in Section 3.4. We detail below the two learning approaches.

1. *Direct optimization.* Direct optimization of the objective function is performed by minimizing the objective function in Eq. (17) directly w.r.t. all parameters $\boldsymbol{\Omega}$ of the model. We denote this approach as **VSLd**. The gradients of the log-likelihood function in the first term on the right side of Eq. (17) are given by:

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

The sum of gradient derivations for H-CRF (for $\nu = 0$) and H-CORF (for $\nu = 1$) can be used to obtain these gradients. The computation of the gradients for the model parameters w.r.t. the regularizers in Eq. (17) is then straightforward. In all our experiments, we used the Limited-memory BFGS method for optimization.

2. *Expectation–Maximization (EM) optimization.* Alternatively, the model parameter can be obtained using the EM algorithm. The EM algorithm [9] is an iterative optimization approach that can be employed to find the latent state parameters $\boldsymbol{\Omega}$ that maximize the VSL-CRF objective (Eq. (17)) in two steps. In the E-step, the posterior probability of the binary latent variable $\nu$ is computed as $P(\nu|x,y)$, i.e., by integrating out the latent states $h$, for each target sequence. Then, the maximum-likelihood parameter estimates of the model parameters $\boldsymbol{\Omega}$ are computed in the M-step. This process is repeated until the convergence of the objective in Eq. (17). More specifically, in the E-step, we compute the posterior probabilities for each target sequence using the auxiliary function:

$$q(\nu_i) = p(\nu_i|\mathbf{y}_i,\mathbf{x}_i,\mathbf{\Omega}^j) \tag{18}$$

This is followed by the M-step, where a new parameter vector $\boldsymbol{\Omega}^{j+1}$ is obtained by maximizing the likelihood function using the current posterior for $\nu$:

$$\begin{aligned}\mathbf{\Omega}^{j+1} = \underset{\Omega}{\operatorname{argmax}} &\sum_{i=1,\ldots,N}\sum_{\nu_i} q(\nu_i)\log P(\mathbf{y}_i,\nu_i|\boldsymbol{x}_i,\mathbf{\Omega}^j) \\ &- \lambda_{n(o)}||\mathbf{\Omega}^j||^2 - \lambda_p\sum_{\nu'}\mathcal{R}_{\nu'}.\end{aligned} \tag{19}$$

In our experiments, we initialized the model with a uniform distribution $q(\nu_i = o) = 0.5$, $q(\nu_i = n) = 0.5$ for all classes, and ran the EM-algorithm until it converged. We denote this learning approach as **VSLem**. It is important to mention that the most important aspect of the **VSLem** approach, compared to the **VSLd**, is that in the latter, the importance of both nominal and ordinal states is equal and does not change during learning. By contrast, through the E-step, the **VSLem** dynamically adapts the weight of each model (nominal vs ordinal) for each sequence. Together with the proposed posterior regularization, this is expected to drive the type of latent states for each sequence to either nominal or ordinal, and thus, avoid over-parametrization of the target data.

*3.3.2.3. Prediction.* Once the model parameters $\boldsymbol{\Omega}$ are learned using either of the proposed approaches (**VSLm**, **VSLd** or **VSLem**), the inference of test data can be performed in two ways, depending on the target task. The first task is sequence-based classification of facial expressions. The goal here is to classify the pre-segmented sequences of facial expressions (e.g., emotions) into one of target classes. In the case of AUs, the goal is to perform detection of the target AU from pre-segmented sequences classified into active (containing activations of the target AU), and 'all other' (containing neutral facial expressions and/or facial expressions of non-target AUs). The assignment of a test sequence to the particular class is accomplished by the MAP rule $y^* = \arg\max_y P(y|\mathbf{x}^*)$. In the case of frame-based classification of target facial expressions, the learned models are used to compute the likelihood of each time-window in the input test sequence. Then, the central frame in the window is assigned the target class, as given by the MAP rule mentioned above.

### 3.4. Posterior regularization

In this section, we show how geometric knowledge of the posterior probability distribution can be used in our optimization framework. This is motivated by recent works [48–50] on posterior regularization in the conditional models, used to improve the parameter learning by incorporating prior knowledge. Formally, let $\Theta$ denote model parameters and $H$ denote hidden variables. Given a set of observed data $\mathcal{D}$, posterior regularization is generally defined as solving a regularized maximum likelihood estimation (MLE) problem:

$$P(y|\mathbf{x}) = \max_{\Theta}\mathcal{L}(\Theta;\mathcal{D}) + \Phi(p(H|\mathcal{D},\Theta)) \tag{20}$$

where $\mathcal{L}(\Theta;\mathcal{D})$ is the marginal likelihood of $\mathcal{D}$, and $\Phi(\cdot)$ is a regularization function of the model posteriors over latent variables. A common definition for $\Phi(\cdot)$ is the KL-divergence between a desired distribution with certain properties over latent variables and the model posterior distribution. In this paper, $H$ corresponds to the

[2] Note that this regularization does not apply to the VSLm approach as the 'hard' selection of the latent states is achieved using the *max* function.

sequence latent state $\nu$. This parameter is not known and no assumptions can be made in order to construct the KL-divergence. However, we make use of the prior knowledge that sequences, which are sampled from the same class should have the same latent states. For instance, we assume that if two sequences $\{\mathbf{y}^1, \mathbf{x}^1\}$ and $\{\mathbf{y}^2, \mathbf{x}^2\}$ are from the same target class $k$, then the conditional probabilities $P(\nu|\mathbf{y}^1 = k, \mathbf{x}^1)$ and $P(\nu|\mathbf{y}^2 = k, \mathbf{x}^2)$ should be similar. Suppose that there are $K$ classes and let $f_{\nu,k}(\mathbf{x}) = P(\nu|\mathbf{y} = k, \mathbf{x})$ be the conditional posterior probability density function for each class defined as $P(\nu|\mathbf{x}, \mathbf{y}) = \sum_h P(h, \nu|\mathbf{x}, \mathbf{y})$. Then, the regularization is performed by minimizing the distance between each element of $f_\nu$ having the same class label. This can be solved by using the graph Laplacian $L$ [51] regularization approach. To this end, we construct a graph $G$ in which each node $n_i$ corresponds to a sequence $\mathbf{x}^i$ with the class label $\mathbf{y}^i$. We connect all nodes with edges $e_{ij}$ that have the weight $s_{ij}$, which is defined by a similarity matrix $S$. In this work, we assign value 1, if and only if $\mathbf{y}^i = \mathbf{y}^j$, $i, j = 1, \ldots, N$, and 0 otherwise. Note, that $y$ is the sequence label and $N$ the total number of sequences. We also do not consider different sequence lengths. This ensures that only the sequences that come from the same class of facial expressions or contain activation of the same AU, are connected. Finally, the graph Laplacian is constructed as $L = D - S$, where $D$ is a diagonal matrix, the entries of which are column-sums of $S$, that is, $D_{ij} = \sum_j S_{ij}$. Then, the proposed posterior regularization $\mathcal{R}_\nu$ is defined as follows:

$$\begin{aligned}\mathcal{R}_\nu &= \frac{1}{2}\sum_{i,j=1}^{m} S_{ij} \cdot (P(\nu|y^i, x^i) - P(\nu|y^j, x^j)) \\ &= \sum_{i=1}^{m} P(\nu|y^i, x^i)^2 D_{ij} - \sum_{i,j=1}^{m} P(\nu|y^i, x^i) P(\nu|y^j, x^j) S_{ij} \\ &= \vec{f}_\nu^T D \vec{f}_\nu - \vec{f}_\nu^T S \vec{f}_\nu \\ &= \vec{f}_\nu^T L \vec{f}_\nu\end{aligned}$$

where

$$\vec{f}_\nu = (P(\nu|y^1, x^1), \ldots, P(\nu|y^m, x^m))^T \tag{21}$$

Note that the larger values of the disparity in $\vec{f}_\nu$ result in a larger regularization loss $\mathcal{R}_\nu$ for state $\nu = \{n, o\}$. The matrix $L$ is positive semi-definite, so $\mathcal{R}_\nu$ is convex in $\vec{f}_\nu$ and by minimizing $\mathcal{R}_\nu$, we get a conditional distribution $f_\nu$ which is sufficiently smooth on the data manifold.

## 4. Experiments

In this section, we evaluate performance of the proposed VSL-CRF model and different learning strategies in the tasks of classification of facial expressions of emotions, and AU detection. The presented experiments are conducted on three publicly available facial expression datasets: Extended Cohn–Kanade (CK+) [45], GEMEP-FERA [46] and DISFA [47]. We also compare the performance of the proposed models with the state-of-the-art methods for both tasks, in the sequence-based classification and frame-based detection settings.

### 4.1. Experimental setup

#### 4.1.1. Datasets

The facial expression datasets used in this work are summarized in Table 2. The **CK+** dataset contains 593 facial expression sequences from 123 different subjects. Each sequence begins with a neutral face and ends at the peak intensity of facial expression of target emotion category. In total, 327 sequences that are labeled in terms of the basic emotions: Anger, Contempt, Disgust, Fear, Happiness, Sadness, or Surprise, are used. We performed 10-fold subject-independent cross-validation on this dataset. The **GEMEP-FERA** dataset contains 87 image sequences of 7 subjects with the per-frame labels for the AU (1, 2, 4, 6, 7, 10, 12, 15, 17, 18, 25 and 26) activations (present or not). Furthermore, in the target videos, each participant shows facial expressions of the emotion categories: Anger, Fear, Joy, Relief or Sadness. We report our results using a 7 fold subject-independent cross validation, where each fold contained image sequences of a different subject. The **DISFA** dataset, contains 32 sequences from 27 subjects. Each sequence in this dataset is 4000 frames long, and each frame is labeled in terms of the intensity level (using FACS) for each AU (1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25 and 26). For our detection approach, we used the frames with the AU intensity higher than 0 as positive examples, and the remaining ones as negative. We performed a 10 fold subject independent cross-validation on this dataset.

**Table 1**
F1-sequence-based results on the **DISFA** database.

| | AU | SVM (SB) | HCRF | HCORF | VSLm | VSLd | VSLem |
|---|---|---|---|---|---|---|---|
| | 1 | 56.1 | 51.4 | 58.3 | 68.9 | 72.3 | **73.7** |
| | 2 | 60.9 | 67.3 | 68.0 | 71.5 | **77.4** | 76.3 |
| Upper | 4 | 61.8 | 63.0 | 57.3 | 68.4 | **72.3** | 66.4 |
| face | 5 | 51.3 | 73.1 | 76.9 | 75.2 | 77.2 | **81.3** |
| | 6 | 68.8 | 70.5 | 64.2 | 74.3 | 72.2 | **74.8** |
| | 9 | 71.4 | 70.3 | 67.7 | 68.5 | **73.5** | 72.2 |
| | 12 | 67.2 | 65.9 | 66.3 | **71.9** | 68.3 | 69.9 |
| | 15 | 52.7 | 61.3 | 56.4 | 64.4 | **68.7** | 68.5 |
| Lower | 17 | 60.5 | 62.4 | 55.3 | 61.2 | 73.4 | **74.3** |
| face | 20 | 57.3 | 61.5 | 57.2 | 63.4 | 71.8 | **73.2** |
| | 25 | 63.8 | 71.2 | 68.4 | **74.2** | 72.3 | 72.4 |
| | 26 | 63.5 | 64.4 | 64.8 | 67.3 | 64.2 | **68.4** |
| | Avg | 61.3 | 65.2 | 62.6 | 69.1 | 72.0 | **72.6** |

Bold data represents the results with the highest performance (highest numbers).

#### 4.1.2. Sequence-based training

The proposed models require sequential data for training and prediction and the CK+ database can be directly used. However, the AU databases GEMEP-FERA and DISFA require a pre-segmentation step in order to extract sequence training data from these databases. We created a training dataset that consists of active and not-active subsequences of each AU. More specifically, from the full dataset, we selected the segments in which the target AU is active (inactive) for the duration of at least 6 frames, and used these as positive (negative) sequences for training. We then balanced the data by removing inactive sequences. Note that we selected the threshold of 6 frames because less than this consistently downgrades the performance on most target AUs, as can be seen from Fig. 6. Once the VSL-CRF models are trained using these pre-segmented data, we apply it in both sequence-based and frame-based manner, as explained in Section 3.3.2.

**Table 2**
F1-frame-based results on the **DISFA** database.

| | AU | SVM (FB) | VSLem | HMTMKL [31] | $l_p$MTMKL [32] | MTFL [55] |
|---|---|---|---|---|---|---|
| | 1 | 53.5 | **75.8** (10) | 72 | 74 | 61 |
| | 2 | 66.8 | 66.2 (6) | 63 | 64 | **70** |
| Upper | 4 | 59.2 | 52.5 (12) | 67 | 68 | **76** |
| face | 5 | **71.8** | 51.7 (6) | 55 | – | – |
| | 6 | 58.8 | 65.3 (12) | 70 | **71** | 65 |
| | 9 | 65.5 | **65.4** (12) | 63 | – | – |
| | 12 | 63.8 | 68.6 (10) | 72 | **76** | – |
| | 15 | 58.3 | **79.6** (10) | 69 | 72 | 68 |
| Lower | 17 | 55.9 | **79.0** (8) | 60 | 63 | 74 |
| face | 20 | 58.3 | 69.5 (8) | 68 | 69 | **71** |
| | 25 | 62.6 | 65.5 (6) | **79** | 74 | – |
| | 26 | 68.7 | **71.5** (10) | 63 | – | – |
| | Avg | 61.9 | 67.6 (*69.1) | 66.8 | **70.1** | 69.3 |

Bold data represents the results with the highest performance (highest numbers).
(*) average F1-Score for the subset of AUs that has been used in $l_p$MTMKL.

| Dataset | Subjects | Videos | Frames/Video | Content | AU annotation | Expression annotation |
|---|---|---|---|---|---|---|
| CK+[45] | 123 | 327 | 20 | posed | binary last frame | per video |
| GEMEP-FERA[46] | 7 | 87 | 20-50 | acted | binary per-frame | - |
| DISFA[47] | 27 | 32 | 4845 | spontaneous | intensity levels per-frame | - |

(a) CK+ [45] (b) GEMEP-FERA [46] (c) DISFA [47]

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

#### 4.1.3. Input features

We used the locations of 49 facial points, extracted from target images sequences using the appearance-based facial tracker in [54]. Fig. 2 depicts the used facial points from each dataset as input features. The pre-processing of the features 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 18, 21, and 24 dimensional feature vectors for the CK+, DISFA and GEMEP-FERA datasets, respectively.

#### 4.1.4. Parameter selection

The model parameters that need to be pre-defined are the fixed number of latent states $C$ and the regularization parametern $\lambda_o, \lambda_n$ and $\lambda_p$. We found the optimal number of latent states by applying a grid search over different settings (in a subject-independent manner). In particular, we applied a two fold cross validation on different AUs from target datasets. To illustrate this, in Fig. 4 we show the F1-scores for sequence-based detection of AU6 from the GEMEP-FERA and DISFA datasets when a different number of latent states is used in the compared models: H-CRF, H-CORF and VSLd. The results drop for the H-CRF model when selecting more than 4 latent states per class. This is mainly because of overfitting but also because of the higher dimensionality of the problem. This effect is not significant for the H-CORF model since the ordinal constrains prevent this model from overfitting. However, in all experiments on all AUs, the F1-measure has a strong increase from 2 to 3 hidden states, which is the number of states corresponding to the temporal phases of expression development (neutral-onset/offset-apex). Adding more states does not improve the models' performance significantly but increases their complexity. Therefore, we set in all our experiments the number of hidden states $C = 3$ for both ordinal and nominal classes. It is important to mention that although VSL-CRF has more latent states per class (3 nominal and 3 ordinal), as noted above, increasing the number of states in H-CRF and H-CORF does not improve their performance significantly. Consequently, the difference in the performance of the compared models (shown in the experiments below) cannot be attributed to the difference in the number of their latent states. Lastly, the regularization parameters $\lambda_{n/o}$ and $\lambda_p$ were set using a grid-search procedure on the validation set found separately for each target fold (no test data were used to perform this validation).

### 4.2. Evaluation measure

We report the classification/detection results using the standard F1-score. This score is widely used for AU-detection and classification of facial expressions of emotions because of its robustness to the imbalance in positive and negative samples, which is very common in the case of AUs. For each AU, the F1-measure is computed based on a frame-based detection (i.e. an AU detection has to be specified for every frame, for every AU, as being either present or absent). We also provide the results for the sequence-based classification, where the F1-score for sequences is computed based on a sequence-based prediction, and then weighted by the number of frames in each sequence. We do so in order to have the fair comparison with the frame-based approaches. We refer to these metrics F1-sequence-based for the sequence based approaches, and the F1-frame-based for the frame-based detection. For emotion classification, we used the F1 score, without weighting with the number of frames in the expression sequence, as methods compared on the CK+ dataset perform the sequence-based classification Fig. 3.

#### 4.2.1. Compared methods

In all our experiments, as the baseline for the classification we also include the results obtained by first applying the multi-class SVMs (with the RBF kernel) and trained/evaluated per frame to obtain the F1-frame-based measure. The sequence labels and the F1-sequence-based measure were obtained by majority voting over the frames within the sequence. The results for H-CRF and H-CORF, were obtained using our own implementation[3] . The initial parameters of the models were set using the same approach as in the VSL-CRF. To compare the performance of target models with the state-of-the-art models for each of target tasks (sequence-based emotion recognition and frame-based AU detection), we report the results from the original papers, as detailed below.

[3] We provide a toolbox with the Matlab code for the compared H-CRF, H-CORF and VSL-CRF models, at http://ibug.doc.ic.ac.uk/resources/DOC-Toolbox/.

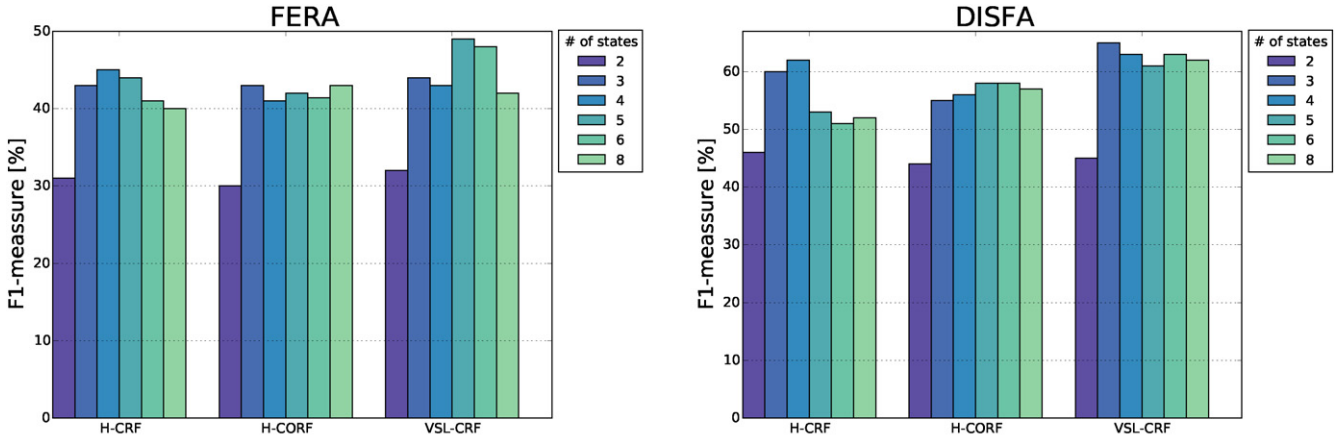
| Emotion | SVM (SB) | HCRF | HCORF | VSLm | VSLd | VSLem | CLM [50] | Cov3D [51] | ITBN [24] | POHCRF [16] | STMexp [8] | TMS [15] | MCSPL [20] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anger | 76.7 | 95.5 | 93.3 | 93.3 | **97.8** | 97.8 | 70.1 | 94.4 | 91.1 | 69.4 | — | 97.9 | 76.3 |
| Contempt | 45.3 | 82.4 | 70.6 | 84.2 | 88.2 | 88.2 | 52.4 | **100.0** | 78.6 | — | — | — | — |
| Disgust | 82.1 | 94.9 | **98.3** | **98.3** | 96.6 | 96.6 | 92.5 | 95.5 | 94.0 | 88.9 | — | 97.9 | 94.1 |
| Fear | 67.4 | 84.0 | 69.2 | **96.0** | 92.0 | **96.0** | 72.1 | 90.0 | 83.3 | 87.7 | — | 90.5 | 86.2 |
| Happy | 86.2 | 95.6 | 97.1 | 97.1 | 97.1 | 98.6 | 94.2 | 96.2 | 89.8 | 98.0 | — | **99.6** | 96.4 |
| Sadness | 62.4 | 64.2 | 79.3 | 87.9 | 87.9 | 87.9 | 45.9 | 70.0 | 76.0 | **97.5** | — | 90.1 | 88.3 |
| Surprise | 87.0 | 98.7 | **100.0** | 98.7 | 97.4 | **100.0** | 93.6 | **100.0** | 91.3 | 98.6 | — | 98.9 | 98.7 |
| Avg | 72.4 | 87.9 | 86.8 | 93.6 | 93.8 | 95.1 (***96.1**) | 74.4 | 92.3 | 86.3 | 90.0 | 94.2 | **95.8** | 90.0 |

**Fig. 3.** Per-sequence classification rate on the CK+ database and comparison with the state-of-the-art.

*4.2.1.1. Sequence-based methods.* Note that some of the methods compared use different number of folds when performing cross-validation on the CK+ dataset. Specifically, PO-HCRF9 (partially observed H-CRF) [18] used a 5-fold cross-validation. In this method, some states are observed during training and represent activations of AUs but the goal is to classify emotions. TMS [17] (Temporal Modeling of Shapes) uses Latent-Dynamic CRFs [25] for a frame-based prediction. However, these predictions are then used to obtain the sequence label. They applied a 4-fold cross validation. ITBN [26] (Interval Temporal Bayesian Network) aims to model temporally overlapping or sequential primitive facial events and the experiments are performed in a 15-fold cross validation setup. Cov3D [53] is based on spatio-temporal covariance descriptors. The descriptors belong to the group of matrices, which can be formulated as a connected manifold. The authors used a 5-fold cross validation. The Constrained Local Method (CLM) [52] is a generic or person-independent face alignment algorithm with goal of finding the shape which is described by a 2D triangulated mesh that fits the target face. They use a 10-fold experimental setup. The MTSL [22] is a multi-task sparse learning framework in which expression recognition and face verification tasks, are coupled to learn specific facial patches for individual expression. Lastly, we compare our method to the state-of-the-art method for target task, STM-ExpLet [13]. The approach combines low-level features from videos with a spatio-temporal manifold learning framework and they evaluate the method using 10-fold cross-validation.

*4.2.1.2. Sequence-based results.* Table 3 shows the results for facial expression recognition from the CK+ dataset. The average classification rate is obtained by unweighted averaging of the results of the 6 basic emotion (*) plus the contempt emotion. Note that while the results of the compared L-CRF models are directly comparable, as they are trained/tested on the same data/folds, this is not the case with the rest of the models as they use different evaluation settings. However, we report their performance for the sake of comparisons. Note also that in this task, i.e., the classification of facial expressions of emotions, the dynamic methods (H-CRF, H-CORF and VSL-CRF) outperform by the large margin the sequence-based SVM classifier that does not account for temporal dynamics. This table also shows that the proposed variable-state method outperforms the other methods that do not have the flexibility to select the best latent states. On the other hand, the proposed VSLem learning strategy improves the classification performance compared to the other two introduced learning methods (VSLm and VSLd). We attribute this to the iterative learning of the latent states, as well as the posterior regularization, which, evidently, together help to increase the discriminative power of the VSL-CRF model. Lastly, the proposed VSLem achieves the state-of-the-art performance in the target task by performing similar or better than the best performing state-of-the-art models, STM-ExpLet and TMS.

Tables 1 and 3 show the results for AU detection on the DISFA and GEMEP-FERA database using pre-segmented sequences. Again, the proposed VSL-CRF model outperforms the models that use only




**(a)** AU6 from the GEMEP-FERA database **(b)** AU6 from the DISFA database

**Fig. 4.** Cross validation over the number of latent states. The tables show the F1-per-sequence measure on AU6 from (a) the GEMEP-FERA and (b) the DISFA datasets w.r.t. the different number of the latent states (nominal and ordinal). In the case of VSL-CRF, the shown number is used separately for nominal and ordinal states.

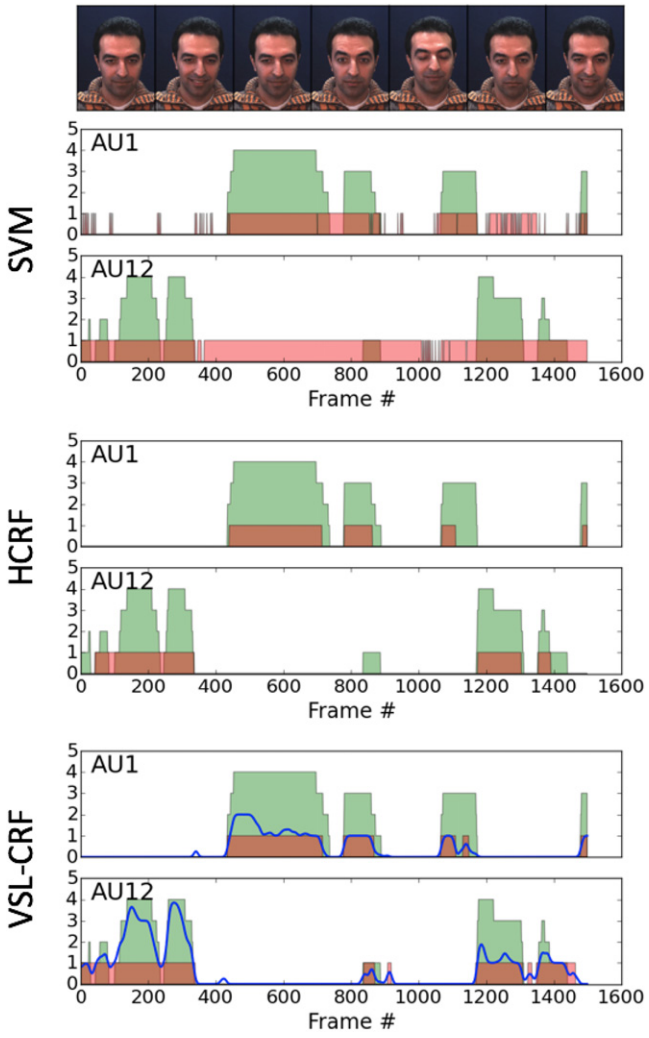



**Fig. 5.** Frame-based predictions of AU1 (Inner Brow Raiser) and AU12 (Lip Corner Puller) from the DISFA database. We used a sliding window with the optimal size around each target frame to obtain the per-frame detection (red box) from the prediction (blue) of the models. The plots also show the annotated 5 intensity levels (green) that have been binarized to train the models for AU detection.

nominal (H-CRF) or ordinal (H-CORF) states, trained/tested on identical data/folds. Furthermore, the highest detection rate is again achieved using the VSLem model on both the DISFA and GEMEP-FERA datasets. Moreover, all the VSL-CRF methods achieve significantly higher results than the other L-CRF models, which is mainly because of the ability to select the optimal states per sequence.

*4.2.1.3. Frame-based methods.* We also compared the variable state models with recent methods for frame-based AU detection. The first related method, Early Fusion (EF) [23], applies a hierarchical Gaussianization and scale-invariant feature transform on motion features. The classification is done by SVMs. In MKL [30], a kernalized SVM is trained for each AU and the outputs are averaged in order to exploit temporal information. CoT (Cascade of Tasks) [39] is trained on sequences and applies segment-based detection. This approach is a combination of three simple algorithms for static-frame-level-detection, segment-level-detection and transition-level detection. Selective Transfer Machine (STM) [28] is based on static SVMs, 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. HMTMKL [31] is a method for multiple AU recognition. A multi-task feature learning (MTFL) algorithm is adopted to learn the shared features among AUs and recognize AUs simultaneously. The AU relations are then modeled by a Bayesian graphical model. Finally, [32] is also a multi task learning approach and applies simultaneous detection of multiple facial AUs by exploiting their inter-relationships.

*4.2.1.4. Frame-based results.* The experiments for per-frame AU detection were performed on the GEMEP-FERA and DISFA database, where we applied a sliding window to each frame in order to obtain the predictions per frame (by assigning the classifier's prediction to the central frame in the window). For each AU, we cross-validated over different window sizes to find the optimal size per AU. The results are shown in Fig. 6. Interestingly, the average window size on the AUs from the GEMEP-FERA dataset is shorter than that of the AUs from the DISFA dataset. This is mainly because both datasets contain facial expressions recorded in different contexts (acted vs. spontaneous), so this difference in the duration of the AU activations is expected. Also, in DISFA, the expressions are less dynamic because the participants respond spontaneously to the watched youtube videos, while in GEMEP-FERA, the participants are actors and show much more dynamic emotions like 'Anger' or 'Fear' with fast facial muscle movements.

Table 4 shows the F1-measure for the detection of each AU from the GEMEP-FERA dataset with the window size reported in brackets. The STM [28], despite the subject adaptation, still fails to reach the full performance of the VSLem model on the mutual set of evaluated AUs. This is attributed to the fact that the STM does not model the temporal dynamics. But again, different settings were used in these evaluations. These results demonstrate again that the assignment of both types of latent states, as done in the VSL-CRF models is critical for achieving superior performance on this task. Table 2 shows the results on the DISFA dataset. The two multi-task learning approaches (MTL) [32, 55] apply simultaneous detection of multiple facial AUs by exploiting their inter-relationships. They also model the correlation among AUs which results in the very high detection rate. The proposed VSL-CRF model reaches the results that are comparable with that of the state-of-the-art. The high F1-frame-based score achieved by both methods demonstrates the importance of both the modeling of the inter-relationships of AUs, as done in the former, and dynamics, as done in the latter. The importance of modelling dynamics for AU detection is demonstrated in Fig. 5. The frame-based detection of AU1 (Inner Brow Raiser) and AU12 (Lip Corner Puller) of the HCRF and VSLem models is smooth and outperform the SVM that is applied in a static manner. Note that the VSLem model succeeds to detect even the segments with very low AU intensity. Examples for AU12 are the segments at frame number 820 and 1450.

*4.2.2. Sequence-based cross-database results*

Detecting AUs across datasets is challenging because of differences in contexts in which this data is recorded (acted vs. spontaneous, illumination, frame rate, etc.). In this experiment, we apply the VSL-CRF models, the H-CRF and H-CORF models, and the baseline SVM on the pre-segmented sequence form the AU databases GEMEP-FERA and DISFA. Tables 6 and 5 show the results for the experiment in which we trained the models using the GEMEP-FERA database and evaluated them on the DISFA database, and the other way round, respectively. We observe that in this setting also the proposed VSL-CRF models outperform nominal- or ordinal-state methods, and the static SVM. This demonstrates the strong generalization capability of the proposed models. It is interesting to note that this difference is much smaller in the results reported in Table 6, where HCRF achieves similar results to VSLem, compared to Table 5, where the HCRF and H-CORF are largely outperformed by the VSL-CRF models. We attribute this to the fact that the acted data (GEMEP-FERA) contains much more variation in facial expressions compared to

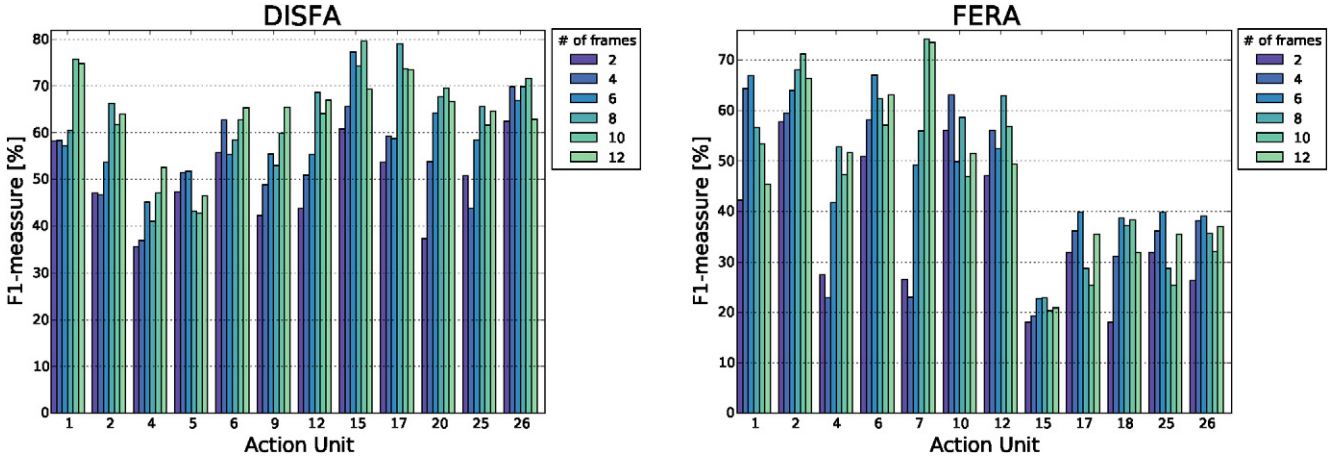



**Fig. 6.** F1-measure per AU for different window sizes for the frame-based VSLem detection.

**Table 3**
F1-sequence-based results on the **GEMEP-FERA** database.

| | AU | SVM (SB) | HCRF | HCORF | VSLm | VSLd | VSLem |
|---|---|---|---|---|---|---|---|
| | 1 | 63.1 | 57.1 | 63.4 | **67.3** | 55.8 | 65.1 |
| | 2 | 62.2 | 65.8 | 64.8 | 63.8 | 64.4 | **71.7** |
| Upper | 4 | 44.7 | 44.4 | 44.2 | 44.2 | **49.7** | 48.2 |
| face | 6 | 57.4 | 53.5 | 51.8 | **58.4** | 53.7 | 54.9 |
| | 7 | 60.3 | 64.2 | 65.4 | 63.2 | 66.2 | **67.5** |
| | 10 | 50.8 | 55.5 | 56.4 | **58.5** | 57.4 | 56.3 |
| | 12 | 54.3 | 45.2 | 43.2 | 53.3 | **54.7** | **54.7** |
| | 15 | 12.4 | 15.3 | 14.9 | 14.4 | 14.2 | **15.5** |
| Lower | 17 | 44.9 | 64.8 | 68.3 | 67.8 | 69.4 | **71.6** |
| face | 18 | 44.0 | 43.1 | 41.7 | **50.3** | 50.1 | 49.8 |
| | 25 | 52.5 | 54.3 | 51.2 | **61.1** | 54.8 | 57.5 |
| | 26 | 48.3 | 33.4 | 35.8 | **49.4** | 44.4 | 48.4 |
| | Avg | 52.0 | 53.5 | 53.4 | 57.7 | 57.2 | **59.0** |

Bold data represents the results with the highest performance (highest numbers).

spontaneous expressions in DISFA dataset. Consequently, the models are learned on more diverse data, allowing them to generalize better to subtle facial expressions, as evidenced by this experiment. We also observe that all three VSL-CRF learning approaches perform similarly in this setting. A possible reason is that since the data distributions vary significantly across the datasets (in terms of number of active examples, as well as the AU co-occurrences), this limits the proposed learning approaches to reach their full performance. Finally, note that the performance on the both datasets drops significantly compared to the results in Tables 1 and 3. For example, for GEMEP-FERA,

**Table 4**
F1-frame-based results on the **GEMEP-FERA** database.

| | AU | SVM (FB) | VSLem | CLM [52] | CoT [39] | STM [28] | MKL [30] | EF [23] |
|---|---|---|---|---|---|---|---|---|
| | 1 | 52.5 | 66.9 (6) | **78** | 64.2 | 68.1 | 61.1 | 57.6 |
| | 2 | 51.8 | 71.2 (10) | **72** | 57.2 | 65.5 | 54.4 | 49.4 |
| Upper | 4 | 42.5 | **52.7** (6) | 43 | 46.6 | 43.3 | 45.4 | 43.6 |
| face | 6 | 55.2 | 67.0 (6) | 66 | **72.9** | 71.6 | 67.0 | 62.3 |
| | 7 | 53.3 | **74.1** (10) | 55 | 67.4 | 66.2 | 65.1 | 61.3 |
| | 10 | 44.9 | 63.0 (4) | 47 | — | — | — | — |
| | 12 | 42.2 | 62.9 (10) | 78 | **78.3** | 82.1 | 75.4 | 71.5 |
| | 15 | 12.2 | 22.9 (10) | 16 | **39.3** | — | — | — |
| Lower | 17 | 31.9 | **50.1** (10) | 47 | 38.6 | 35.9 | 36.7 | 30.1 |
| face | 18 | 42.4 | 38.7 (6) | **45** | — | — | — | — |
| | 25 | **41.3** | 39.7 (6) | 31 | — | — | — | — |
| | 26 | 49.5 | 39.0 (6) | **54** | — | — | — | — |
| | Avg | 46.5 | 58.7 (*__63.6__) | 56.0 | 57.1 | **61.8** | 57.9 | 57.6 |

Bold data represents the results with the highest performance (highest numbers).
(*) average F1-Score for the subset of AUs that has been used in $l_p$MTMKL.

**Table 5**
Per-sequence classification rate on the cross dataset experiment DISFA →**GEMEP-FERA**.

| AU | SVM | HCRF | HCORF | VSLm | VSLd | VSLem |
|---|---|---|---|---|---|---|
| 1 | 40.0 | **44.2** | 39.7 | 42.6 | 43.3 | 43.7 |
| 2 | 40.4 | 44.8 | **47.4** | 43.3 | 42.8 | 41.2 |
| 4 | 33.3 | 33.5 | 25.3 | 22.4 | **34.8** | 34.0 |
| 6 | 57.7 | 54.1 | 46.3 | **58.7** | 49.7 | 54.5 |
| 12 | 23.7 | 35.9 | 34.5 | 33.0 | 36.0 | **37.4** |
| 17 | 22.2 | 29.4 | 16.9 | 19.8 | 24.6 | **25.6** |
| 25 | 37.2 | **67.9** | 44.6 | 46.7 | 44.4 | 45.6 |
| 26 | **37.5** | 31.4 | 36.0 | 37.3 | 33.2 | 32.4 |
| AVG | 35.3 | 38.9 | 36.3 | 37.9 | 38.6 | **39.2** |

Bold data represents the results with the highest performance (highest numbers).

the results on the used set of AUs from 60.2% to 39.2% for the best performing model. This indicates the importance of accounting for the dataset-differences during modeling of facial expressions.

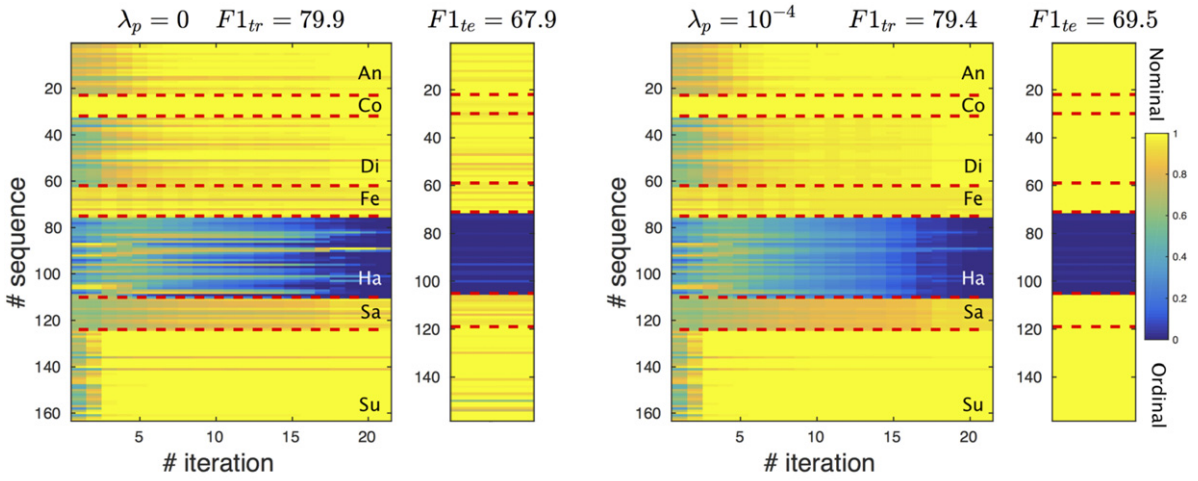
*4.2.3. The effect of posterior regularization*

On all datasets, the VSLd and VSLem outperforms VSLm. This is mainly attributed to the more flexible representation of the latent states as well as the additional posterior regularization. To get some insights into the behavior of the posterior regularization during the learning process, we performed additional experiment on the CK+ dataset. Specifically, we trained the VSLem model with and without the posterior regularization and monitored the parameter for each EM-iteration (the graphs showing the changes in the nominal/ordinal states on the training data). The training/test sets consisted of 162 sequences each, and are sorted according to the sequence label. The results are shown in Fig. 7. The bar on the right side of each main figure shows the contribution of ordinal/nominal states for the prediction of the test sequences. We can see that the

**Table 6**
Per-sequence classification rate on the cross dataset experiment GEMEP-FERA →**DISFA**.

| AU | SVM | HCRF | HCORF | VSLm | VSLd | VSLem |
|---|---|---|---|---|---|---|
| 1 | 28.5 | 32.8 | 29.6 | 35.2 | **40.0** | 35.2 |
| 2 | 37.2 | 45.7 | 41.3 | **49.4** | 49.3 | 40.9 |
| 4 | 25.9 | 44.9 | 29.2 | 24.8 | 35.5 | **40.9** |
| 6 | **50.8** | 44.6 | 39.9 | 48.6 | 42.6 | 48.4 |
| 12 | 21.2 | 32.1 | 26.2 | **42.7** | 28.5 | 39.6 |
| 17 | 26.6 | 23.1 | 21.9 | 25.4 | 25.3 | **32.7** |
| 25 | 42.1 | 46.5 | 50.5 | 52.6 | **53.2** | 45.3 |
| 26 | 22.2 | 34.1 | 33.0 | 33.3 | 37.8 | **38.0** |
| AVG | 34.3 | 36.7 | 33.9 | 39.0 | 39.1 | **40.1** |

Bold data represents the results with the highest performance (highest numbers).




**(a)** Posterior regularization not active **(b)** Posterior regularization active

**Fig. 7.** The visualization of the learning of the latent variable $\nu = \{ordinal, nominal\}$ within the VSLem model applied to the CK+ database. The evaluation was performed using 2-fold subject-independent evaluation using equal number of training/test data. The posterior probabilities $P(\nu|y,x)$ are shown for each sequence after the maximization step of each EM-iteration. The plots in (a) and (b) shows the learning process without and with the posterior regularization (using the optimal valdiation parameter $\lambda_p$).

emotion happiness exhibited a strong ordinal structure as encoded with its ordinal states, while the other emotion were predicted using the nominal states. The figure on the right shows the same learning process with active posterior regularization. Again, the emotion happiness was trained and predicted using mainly the ordinal states but all other emotions mainly preferred using the nominal states during training and inference, as the result of the regularization. The learned type of the latent states is also consistent on the test data. Finally, although only emotion happiness showed strong ordinal nature, as learned from the employed features of facial expressions, the nominal states selected for the other emotion categories do not imply that there is no ordinal structure in their facial expressions but that the nominal states were a better fit for the target data used in this experiment. Note also that when the posterior regularization is used, the F1-sequence-based measure on the test sets is higher (69.5% vs. 67.9%), demonstrating the benefit of the posterior regularization. Furthermore, note that this regularization enforces the model to converge to either nominal or ordinal states during the model learning.

## 5. Conclusions

In this paper, we proposed a novel variable-state 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, the proposed model 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). For this model, we proposed two novel learning strategies and the posterior regularization of the latent states, resulting in a more robust model for the target tasks. This leads to superior performance compared to traditional latent CRF models. We also showed on three facial expression datasets that the proposed model performs similar or better than the state-of-the-art for the task of sequence-based facial expression recognition, and that it reaches state-of-the-art performance for the task of per-frame AU-detection. The future work should focus on more detailed analysis of the learning of the target latent states within each emotion class and AU (e.g., the automated selection of the window size for each AU), as well as analysis of the relations between the learned latent states and the temporal aspects of facial expressions such as their temporal phases and intensity. Also, extending the proposed approach so that it can handle simultaneous detection of multiple AUs, and its adaptation to previously unseen datasets, are also interesting avenues to pursue.

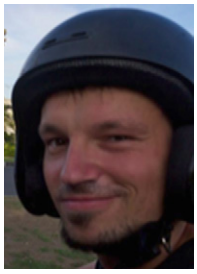
## Acknowledgments

This work has been funded by the European Community Horizon 2020 [H2020/2014–2020] under grant agreement no. 645094 (SEWA). The work of Vladimir Pavlovic has been funded by the National Science Foundation under grant no. IIS0916812.

## References

[1] Maja Pantic, Machine analysis of facial behaviour: naturalistic and dynamic behaviour, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 364 (1535) (2009) 3505–3513.
[2] O. Rudovic, V. Pavlovic, M. Pantic, Context-sensitive dynamic ordinal regression for intensity estimation of facial action units, TPAMI 37 (5) (2014) 944–958.
[3] L .J.M. Rothkrantz, M. Pantic, Automatic analysis of facial expressions the state of the art, TPAMI 22 (12) (2000) 1424–1445.
[4] Evangelos. Sariyanidi, Hatice. Gunes, Andrea. Cavallaro, Automatic analysis of facial affect a survey of registration, representation, and recognition, TPAMI 37 (6) (2015) 1113–1133.
[5] J.F. Cohn, P. Ekman, Measuring facial actions. In the new handbook of methods in nonverbal behavior research, 2005. pp. 9–64.
[6] P. Ekman, W.V. Friesen, J.C. Hager, Facial action coding system (FACS): manual, A Human Face (2002)
[7] M. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, Transactions on Systems, Man, and Cybernetics 42 (1) (2012) 28–43.
[8] Z. Zeng, M. Pantic, G. Roisman, T.S. Huang, A survey of affect recognition methods audio, visual, and spontaneous expressions, TPAMI 31 (1) (2009) 39–58.
[9] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), 2006.
[10] J.D. Lafferty, A. McCallum, F.P.e.r.e.i.r.a. fields, Conditional random probabilistic models for segmenting and labeling sequence data, ICML (2001) 282–289.
[11] M. Kim, V. Pavlovic, Hidden conditional ordinal random fields for sequence classification, Machine Learning and Knowledge Discovery in Databases 6322 (2010) 51–65.
[12] O. Rudovic, V. Pavlovic, M. Pantic, Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation, CVPR (2012) 2634–2641.
[13] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition pages 1749-1756, 2013.
[14] S. Wang, A. Quattoni, L.P. Morency, D. Demirdjian, D. Trevor, Hidden conditional random fields for gesture recognition, CVPR (2006) 1097–1104.
[15] C. Sminchisescu, A. Kanaujia, D. Metaxas, Conditional models for contextual human motion recognition, CVIU 104 (2) (2006) 210–220.

[16] A. Quattoni, S. Wang, L. Morency, M. Collins, T. Darrell, Hidden conditional random fields, TPAMI 29 (10) (2007) 1848.
[17] S. Jain, Hu. Changbo, J.K. Aggarwal, Facial Expression Recognition with Temporal Modeling of Shapes., 2011, 1642–1649.
[18] K. Chang, T. Liu, S. Lai, Learning Partially-observed Hidden Conditional Random Fields for Facial Expression Recognition, 2009, 533–540.
[19] C. Shan, S. Gong, P.W. McOwan, Conditional Mutual Infomation Based Boosting for Facial Expression Recognition, 2005.
[20] G . Littlewort, M.S. Bartlett, I. Fasel, J. Movellan, Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction, 2003.
[21] M. Pantic, L. Rothkrantz, Facial action recognition for facial expression analysis from static face images, Systems, Man, and Cybernetics, Part B: Cybernetics 34 (3) (2004) 1449–1461.
[22] Metaxas. D.N., L. Zhong, P. Liu, J.H.u.a.n.g. Yang, Learning multiscale active facial patches for expression analysis, Transactions on Systems, Man, and Cybernetics, Part B 45 (8) (2015) 1499–1510.
[23] T. Usman, K.L. Hsiang, L. Zhen, Z. Xi, W. Zhaowen, L. Vuong, S.H. Thomas, X.L. Tony, X. Han, Emotion Recognition from an Ensemble of Features, 2011, 872–877.
[24] L. Shang, K. Chan, Nonparametric Discriminant HMM and Application to Facial Expression Recognition, 2009, 2090–2096.
[25] N. Sebe, M. Lew, Y. Sun, I. Cohen, T. Gevers, T. Huang, Authentic facial expression analysis, Image and Vision Computing 25 (12) (2007) 1856–1863.
[26] Z. Wang, S. Wang, Q. Ji, Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition, 2013, 3422–3429.
[27] M.F. Valstar, I. Patras, M. Pantic, Facial Action Unit Detection Using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data, 2005.
[28] W. Chu, F. Torre, J. Cohn, Selective transfer machine for personalized facial action unit detection, CVPR (2013) 3515–3522.
[29] S. Chew, S. Lucey, P. Lucey, S. Sridharan, J.F. Cohn, Improved Facial Expression Recognition via Uni-hyperplane Classification., 2012, 2554–2561.
[30] T. Sénéchal, V. Rapp, H. Salam, R. Seguier, K. Bailly, L. Prevost, Facial action recognition combining heterogeneous features via multikernel learning 42 (4) (2012) 993–1005.
[31] Y. Zhu, S. Wang, L. Yue, Q. Ji, Multiple-facial Action Unit Recognition by Shared Feature Learning and Semantic Relation Modeling, 2014, 1663–1668.
[32] Cohn. J. F, X. Zhang, M.H. Mahoor, S.M. Mavadati, A lp-norm MTMKL Framework for Simultaneous Detection of Multiple Facial Action Units., 2014, 1104–1111.
[33] Y. Song, D. McDuff, D. Vasisht, A. Kapoor, Exploiting Sparsity and Co-occurrence Structure for Action Unit Recognition, 2015.
[34] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, TPAMI (11) (2010) 1940–1954.
[35] B. Jiang, M. Valstar, M. Pantic, Facial Action Detection Using Block-based Pyramid Appearance Descriptors, 2012.
[36] L. van Maaten, E. Hendriks, Action unit classification using active appearance models and conditional random fields, Cognitive Processing 13 (2) (2012) 507–518.
[37] B. Jiang, M. Valstar, B. Martinez, M. Pantic, A dynamic appearance descriptor approach to facial actions temporal modeling, Transactions on Cybernetics 44 (2) (2014) 161–174.
[38] O. Rudovic, V. Pavlovic, M. Pantic, Kernel conditional ordinal random fields for temporal segmentation of facial action units, ECCV'W (2012)
[39] X. Ding, W. Chu, F. De la Torre, J. Cohn, Q. Wang, Facial Action Unit Event Detection by Cascade of Tasks, 2013, 2400–2407.
[40] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, ICML (2001) 282–289.
[41] P. McCullagh, Regression models for ordinal data, Journal of the Royal Statistical Society, Series B 42 (1980) 109–142.
[42] M. Kim, V. Pavlovic, Structured output ordinal regression for dynamic facial emotion intensity prediction, ECCV (2010) 649–662.
[43] A. Quattoni, M. Collins, T. Darrell, Conditional random fields for object recognition, NIPS (2004) 1097–1104.
[44] M. Held, P. Wolfe, H. Crowder, Validation of subgradient optimization, Mathematical programming 6 (1) (1974) 62–88.
[45] Saragih, P. Lucey, F. Jeffrey, T. Cohn, J. Kanade, Z. Ambadar, I. Matthews, The Extended Cohn–Kanade Dataset (Ck+): A Complete Dataset for Action Unit and Emotion Expression, 2010.
[46] M. Valstar, B. Jiang, M. Mehu, M. Pantic, K. Scherer, The First Facial Expression Recognition and Analysis Challenge, 2011, 921–926.
[47] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J.F. Cohn, DISFA: a spontaneous facial action intensity database, Transactions on Affective Computing 4 (2) (2013) 151–160.
[48] K. Ganchev, J. Graça, J. Gillenwater, B. Taskar, Posterior regularization for structured latent variable models, The Journal of Machine Learning Research 11 (2010) 2001–2049.
[49] J. Zhu, N. Chen, E.P. Xing, Bayesian inference with posterior regularization and applications to infinite latent SVMs, The Journal of Machine Learning Research 15 (1) (2014) 1799–1847.
[50] K. Ganchev, D. Das, Cross-lingual Discriminative Learning of Sequence Models with Posterior Regularization, 2013, 1996–2006.
[51] F.R. Chung, Spectral graph theory (CBMS regional conference series in mathematics, no. 92), American Mathematical Society 5 (2). (1997) 5-6.
[52] S.W. Chew, P. Lucey, S. Lucey, J. Saragih, J.F. Cohn, S. Sridharan, Person-independent Facial Expression Detection Using Constrained Local Models, 2011, 915–920.
[53] A. Sanin, C. Sanderson, M. Tafazzoli Harandi, C.L. Brian, Spatio-temporal Covariance Descriptors for Action and Gesture Recognition, 2013, 103–110.
[54] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Incremental face alignment in the wild, CVPR (2013)
[55] X. Zhang, M. Mahoor, R.D. Nielsen, On Multi-task Learning for Facial Action Unit Detection, 2013, 202–207.



**Robert Walecki** received his MSc degree in Physics from the The Ruprecht-Karls-University in Heidelberg, Germany, in 2013. He is currently working towards his Ph.D. degree at the Department of Computing, Imperial College London, London, UK. His research interests span the areas of Computer Vision, Pattern Recognition, Machine Learning and, in particular, human-computer interaction and automatic human behavior analysis.

**Ognjen Rudovic** received his BSc degree in Automatic Control from Faculty of Electrical Engineering, University of Belgrade, Serbia, in 2007, MSc degree in Computer Vision from Computer Vision Center (CVC), Universitat Autonoma de Barcelona, Spain, in 2008, and PhD in Computer Science from Imperial College London, UK. He is currently a Research Fellow at the Computing Department, Imperial College London, UK. His research interests are in automatic recognition of human affect, machine learning and computer vision.

**Vladimir Pavlovic** received the PhD degree in electrical engineering from the University of Illinois at Urbana-Champaign in 1999. From 1999 until 2001, he was a member of the research staff at the Cambridge Research Laboratory, Massachusetts. He is an associate professor in the Computer Science Department at Rutgers University, New Jersey. Before joining Rutgers in 2002, he held a research professor position in the Bioinformatics Program at Boston University. His research interests include probabilistic system modeling, time-series analysis, computer vision, and bioinformatics.

**Maja Pantic** is Professor in Affective and Behavioural Computing at Imperial College London, Computing Dept., UK, and at the University of Twente, Dept. of Computer Science, Netherlands. She received various awards for her work on automatic analysis of human behaviour including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She currently serves as the Editor in Chief of Image and Vision Computing Journal, and as an Associate Editor for IEEE Trans. on Systems, Man, and Cybernetics Part B and IEEE TPAMI.