Image and Vision Computing xxx (2016) xxx–xxx

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 xxxx

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/© 2016 Elsevier B.V. All rights reserved.

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