Sequential Emotion Recognition using Latent-Dynamic Conditional Neural Fields

Julien-Charles Lévesque Louis-Philippe Morency Christian Gagné

Abstract— **A wide number of problems in face and gesture analysis involve the labeling of temporal sequences. In this paper, we introduce a discriminative model for such sequence labeling tasks. This model involves two layers of latent dynamics, each with their separate roles. The first layer, the neural network or gating layer, aims to extract non-linear relationships between input data and output labels. The second layer, the hidden-states layer, aims to model temporal sub-structure in the sequence by learning hidden-states and their transition dynamics. A new regularization term is proposed for the training of this model, encouraging diversity between hidden-states. We evaluate the performance of this model on an audiovisual dataset of emotion recognition and compare it against other popular methods for sequence labeling.**

I. Introduction

Identifying activities in unsegmented video sequences is a frequently encountered problem in face and gesture analysis, but also in computer vision in general. Sequence labeling methods try to solve this problem by learning a mapping between the sequence of input features (e.g., audiovisual signals from video sequences) and the sequence of output labels (e.g., behaviors or emotions expressed in the video). When this mapping between input features and output labels is linear, sequential discriminative models such as Conditional Random Fields (CRFs) have shown great performance [1], [2], often outperforming their generative counterpart.

Two of the main challenges for sequential labelling problem such as sequential emotion recognition are (1) how to deal with complex non-linear input features, and (2) how to model important sub-structure in label sequence. Facial expression recognition needs to integrate information from multiple cues (eye brows, eyes, mouth, cheeks) and these expressions often have multiple phases (onset, peak, offset). The same thing is true for body gestures where multiple part of the body are used to perform a gesture [3] and these gestures have phases (again: onset, peak, offset).

In this paper, we introduce the Latent-Dynamic Conditional Neural Fields (LDCNF) to solve the problem of unsegmented sequence labeling with non-linear input features and label sub-structure. This new graphical model is able to capture the hidden sub-structure of a class sequence and at the same time learn the non-linear relationships of complex input features and class labels. As shown in Fig. 1, the LDCNF model contains two layers of latent variables. Each layer serves a different purpose. The first layer is known as the neural network or gate layer and aims to learn the non-linear mapping from the input sequences. The second layer is known as the dynamic layer and models temporal sub-structure from the class sequence.

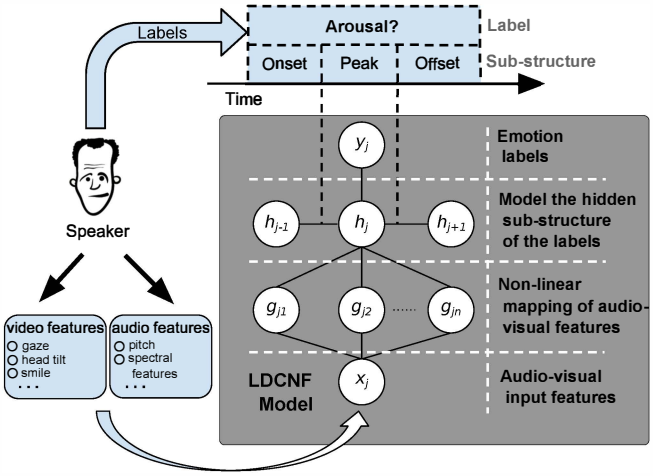

Fig. 1. Overview of the LDCNF model and the problem of emotion recognition from audiovisual data. The grey box represents the LDCNF model and placed around it are the main elements of the problem. The index j used in the LDCNF box represents a time frame.

A key novelty of our LDCNF model is its objective function which include a new regularization term specifically designed to reduce the potential overlap between the two layers of latent variables. This is an important issue that often arises when a probabilistic model contains more than one layer of latent variables. Both layers could end up learning the same things, if they are unconstrained in their tasks. The new regularization term rewards diversity between latent variables of different layers by penalizing cases where the weights for different hidden states are similar.

We evaluate our LDCNF model using the publicly available dataset recently released as part of the Audio-Visual Emotion Challenge (AVEC2011) [4]. This dataset contains 63 unsegmented sequences of natural interactions. This dataset involves complex relationship between the multimodal input modalities (e.g., audio and video) and the emotional labels. Finally, we compare our LDCNF model

J.-C. Lévesque and C. Gagné are with the Laboratoire de vision et systèmes numériques, Université Laval, Québec, QC, Canada julien-charles.levesque.1@ulaval.ca, christian.gagne@gel.ulaval.ca

L.-P. Morency is with the Institute for Creative Technologies, University of Southern California, Los Angeles, CA 90094, USA morency@ict.usc.edu

This material is based upon work supported by the National Science Foundation under Grant No. 1118018 and the U.S. Army Research, Development, and Engineering Command. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Government.

with other popular approaches for unsegmented activity recognition.

The following section describes related work on sequence labeling and audio-visual emotion recognition. The Section III gives some background information about the LDCRF model. Section IV presents our new Latent-Dynamic Conditional Neural Field model and describes our learning approach including the new regularization term. Sections V and VI gives a detailed description of our experiments and discusses our results. We conclude with Section VII.

II. Related Work

Extensions of the CRF model have been proposed to better model the natural sub-structure happening in many sequence labeling tasks such as emotion recognition. Two such examples are the Hidden Conditional Random Field (HCRF) for *segmented* sequences by Quattoni et al. [5] and the Latent-Dynamic Conditional Random Field (LDCRF) for unsegmented sequences by Morency et al. [6]. Both models incorporate hidden state variables which model the substructure of a class sequence and, in the case of the LDCRF, learn dynamics between class labels. Although they succeed in learning the substructure in gestures or activities, this family of CRF models have a harder time to learn complex non-linear relationships.

The recently introduced Conditional Neural Field (CNF) proposed to address this issue by adding a hidden layer to the CRF model which contains gate functions, each acting as a local neuron or feature extractor [7]. The CNF model can automatically learn an implicit nonlinear representation of features and can capture more complicated relationships between the inputs and outputs. A key advantage of the CNF model is that it can learn these non-linear relationships while keeping the learning and inference procedures efficient using a dynamic programming algorithm. While good results have been shown on protein secondary structure prediction and handwriting recognition, the CNF does not explicitly model the sub-structure of the class sequence which, as we show in our experiments, is important for unsegmented activity recognition.

Van der Maaten et al. [8] used a similar intuition and introduced the Hidden-Unit Conditional Random Field, where they added a layer of binary neurons. This layer outputs a binary representation of the input data and also serves to extract non-linear relationships between the input features and output labels. It was tested on optical character recognition, sentence labeling, part-of-speech tagging, and protein secondary sub-structure prediction. Shyr et al. [9] proposed a kernel method for sequence dimension reduction which also fares well compared to the previous literature in dimensionality reduction, but their method was only applied on segmented sequences. None of these models contain two layers of latent variables to model both the label sub-structure and the non-linearity between input features.

A. Emotion recognition

Papers [10] and [11] include surveys on emotion recognition to which are refered readers new to the field. Of interest is the work by Nicolaou et al. [12], who ran experiments on the classification of spontaneous affect based on Audio-Visual features using coupled Hidden-Markov Models. They showed that using the likelihoods produced from separate HMMs as inputs to other classifiers can be beneficial. Wollmer et al. [13] used Conditional Random Fields (CRF) for discrete emotion recognition based on a selection of acoustic features. In addition, they use Long Short-Term Memory Recurrent Neural Networks to perform regression analysis on these two dimensions. Both of these approaches demonstrate the benefits of including temporal information when approaching emotion recognition in dimensional space.

Eyben et al. [14] fused different visual and audio modalities in order to analyze human affect in valence and expectation dimensions. They found that high level event-based features such as smiles, head nods and laughter were better suited for their task than low level signal-based features such as facial feature points and spectral information.

Ramirez et al. [15] used LDCRFs to recognize presence of emotions in audio, visual and audiovisual signals for the AVEC 2011 challenge. By using high level features, they were able to produce the best results for the visual sub-challenge. Our experiments present a comparison of the performance of our model with the LDCRF model using the same input features.

Jain et al. [16] applied LDCRFs to model the temporal dynamics of face shapes for emotion recognition and showed an improvement in performance compared to using only facial appearance. Rudovic et al. [17] also used an extension of HCRFs, hidden conditional ordinal random fields (H-CORF), for expression recognition in learned manifolds.

III. Latent-Dynamic Conditional Random Fields

LDCRFs [6] were designed to learn the sub-structure in sequence labels. The goal is to learn a mapping between a sequence of observations $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$ (e.g., features provided by facial trackers and audio feature extractors) and a sequence of labels $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ (e.g., presence of an emotion or not). Each label y_j is contained in the set of all possible labels $y_j \in \mathcal{Y}$, and each observation is a feature vector $x_j \in \mathbb{R}^d$. For each sequence, a series of hidden variables serve to model the hidden or latent dynamics of the process, $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$, each part of a finite set of hidden states $h_j \in \mathcal{H}$. The latent conditional model is represented as :

$$P(y|x,\theta) = \sum_h P(y|h,x,\theta) \cdot P(h|x,\theta), \qquad (1)$$

where θ is the parameter vector defining the model. To keep training and inference tractable, the sets of hidden states are forced to be disjoint. Each h_j is a member of a set $\mathcal{H}_{y_j}$ of possible hidden states for the class label y_j, thus sequences which have any $h_j \notin \mathcal{H}_{y_j}$ will have $P(y|x,\theta) = 0$.

The model is then expressed as:

$$P(y|x,\theta) = \sum_{h:\forall h_j \in \mathcal{H}_{y_j}} P(h|x,\theta), \quad (2)$$

$$P(h|x,\theta) = \frac{1}{Z(x,\theta)} \exp\left(\sum_k \theta_k \cdot F_k(h,x)\right), \quad (3)$$

where the partition function Z is defined as:

$$Z(x,\theta) = \sum_h \exp\left(\sum_k \theta_k \cdot F_k(h,x)\right). \quad (4)$$

F_k is defined as:

$$F_k(h,x) = \sum_{j=1}^{m} f_k(h_{j-1}, h_j, x, j), \quad (5)$$

and each feature function $f_k(h_{j-1}, h_j, x, j)$ is either a vertex function $v_{h,f}(h_j,\ x, j)$ or an edge function $e_{h,h'}(h_{j-1}, h_j, x, j)$. The first depend only on neighboring observations in the sequence while the second depends on adjacent hidden variables in the sequence and models transitions between hidden states. The feature functions take the following forms :

$$v_{h,f}(h_j, x, j) = \delta[h_j = h] \cdot x_{jf}, \quad (6)$$

$$e_{h,h'}(h_j, x, j) = \delta[h_j = h] \cdot \delta[h_{j-1} = h'], \quad (7)$$

where $\delta[h_j = h]$ is an indicator function, equal to one only if the hidden state at position j is h.

LDCRFs were used extensively for gesture recognition with a small number of dimensions, but fall short on tasks requiring the use of a high number of continuous features. More recent approaches like conditional neural fields handle better this feature complexity, but do not explicitly model the hidden label sub-structure. In the next section, we will study how to take advantage of such approaches for the LDCNF model.

IV. Latent-Dynamic Conditional Neural Fields

We define our latent-dynamic conditional neural field model by adding a single-layer neural network as a preprocessing layer to the LDCRF model (see Fig. 2). This provides a better representation of the input data and helps with the modeling of the hidden dynamics.

For this model, Equations 3 and 5 from the previous section remain identical, but we will modify the vertex feature functions so that they include a single-layer neural network. The new vertex feature functions thus take the following form :

$$v_{h,g}(h_j, x, j) = \text{gate}(\theta_g^G \cdot x_j) \cdot \delta[h_j = h], \quad (8)$$

where θ_g^G is a vector of weights for the gate g, $\text{gate}(\cdot)$ is a gating function, in this work the logistic function, $\text{gate}(x) = 1/(1+\exp(-x))$. The parameter vector is split in three sub-vectors, one for each type of feature function, respectively edge, vertex, and gate functions, giving $\theta = [\theta^E, \theta^V, \theta^G]$.

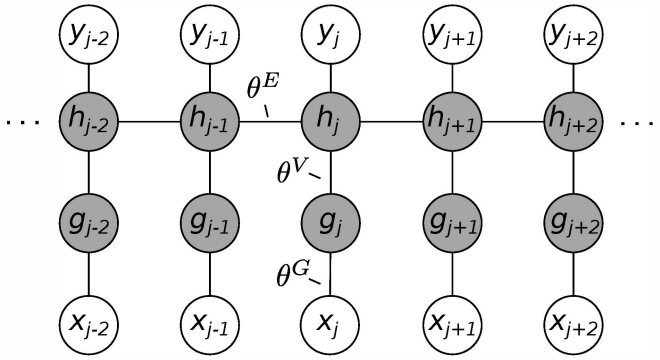

Fig. 2. The Latent-Dynamic Conditional Neural Field model.

This model has two layers of hidden dynamics, the gating layer and the hidden states layer. As with neural networks, it is not always clear whether an additional layer will help generalization. Both layers could end up learning the same things, since they are unconstrained in their tasks. In the next section, we will discuss strategies to constrain learning in a way that will make both layers useful.

A. Learning parameters

This model is trained by log-likelihood maximization with gradient ascent. In this work, the LBFGS method was used [18] because of its speed and robustness, but other methods could be suitable. Given a training set of n labelled sequences (X_i, Y_i), the objective function is as follows:

$$L(\theta) = \sum_{i=1}^{n} \log P(Y_i|X_i,\theta) - \frac{1}{2\sigma^2}||\theta||^2 - \alpha R_{hg}. \quad (9)$$

The first term of the previous equation is the log-likelihood of each individual sequence with the current model. The second term is the log of a Gaussian prior with variance σ^2, i.e., $P(\theta) \sim \exp(\frac{1}{2\sigma^2}\|\theta\|^2)$. The last term is a new regularization term aiming to constrain the training of the weights between the hidden states layer and the gates layer. This term will be high if the weights are similar for different hidden states, and low otherwise. More precisely, it is the sum of dot products between weight vector associated with each hidden state :

$$R_{hg} = \sum_{j=1}^{m} \sum_{k=j+1}^{m} \theta_{h_j}^V \cdot \theta_{h_k}^V \quad (10)$$

This regularization term will encourage a diversity between hidden states. It will also reduce the probability that both layers model the same dynamics. The α parameter allows to control the strength of this regularization.

The log-likelihood of a single training sequence X_i, Y_i is given by:

$$\log P(Y_i|X_i,\theta) = \log \sum_{h \in \mathcal{H}_{Y_i}} P(h|x,\theta) \quad (11)$$

$$\log P(Y_i|X_i,\theta) = \log \sum_{h \in \mathcal{H}_{Y_i}} \exp\left(\sum_k \theta_k \cdot F_k(h,x)\right) - \log Z(x,\theta). \quad (12)$$

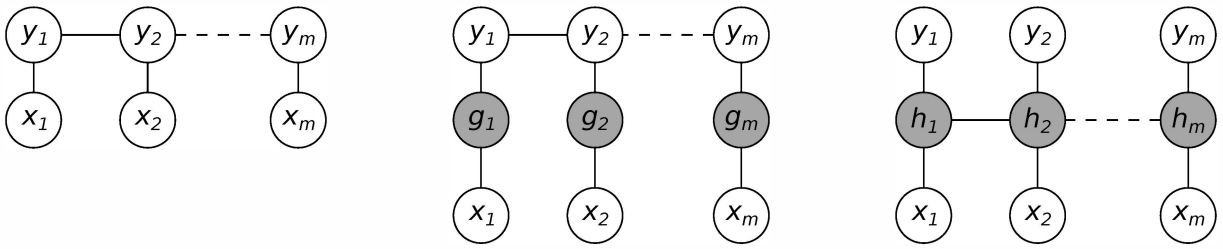

Fig. 3. Comparison of CRF, CNF and LDCRF models (from left to right).

The derivative of this log-likelihood with respect to an arbitrary parameter θ_d is:

$$\frac{\partial L_i}{\partial \theta_d} = \sum_h P(h|y_i, x_i, \theta) \cdot \frac{\partial \left(\sum_k \theta_k \cdot F_k(h, x)\right)}{\partial \theta_d} - \sum_{y',h} P(y', h|y_i, x_i, \theta) \cdot \frac{\partial \left(\sum_k \theta_k \cdot F_k(h, x)\right)}{\partial \theta_d}. \tag{13}$$

According to the three types of parameters in the model, Equation 13 will take three different forms. The gradients for edge and vertex features are the same as for classical LDCRFs, while the gradients for gate features take the following form:

$$\frac{\partial L_i}{\partial \theta^G_{g,f}} = \sum_h P(h|y_i, x_i, \theta) \cdot \theta^V_{h,g} \cdot \frac{\partial \left(\sum_{j=1}^m v_{h,g}(h_j, x, j)\right)}{\partial \theta^G_{g,f}} - \sum_{y',h} P(y', h|y_i, x_i, \theta) \cdot \theta^V_{h,g} \cdot \frac{\partial \left(\sum_{j=1}^m v_{h,g}(h_j, x, j)\right)}{\partial \theta^G_{g,f}}, \tag{14}$$

where

$$\frac{\partial v_{h,g}(h_j, x, j)}{\partial \theta^G_{g,f}} = \text{gate}(\theta^G_g \cdot x_j) \cdot (1 - \text{gate}(\theta^G_g \cdot x_j)) \cdot \delta[h_j = h]. \tag{15}$$

Using the forward-backward algorithm [19], the gradient can be computed efficiently. Similarly to CNFs and LDCRFs, the training of an LDCNF model is a non convex optimization problem.

V. Experiments

We analyse the performance of the LDCNF model on a multimodal dataset for the recognition of emotions, the Audio/Visual Emotion Challenge 2011 (AVEC2011) [4], [20]. Twenty participants were recorded while holding conversations with an operator who adopted in sequence roles designed to evoke emotions in the participants, producing a total of 63 sequences. The presence of emotion was first labelled on a continuous scale or zero to one, then the final labels were produced by thresholding these degrees of emotion. In this experiment, we aim at recognizing the emotion of arousal.

The video data consists of a 780 x 580 pixel resolution video recorded at 49.979 frames per second, with one label per frame (a screenshot is provided in Fig. 4). The audio data was recorded at 48 kHz with 24 bits per sample, with one label per word provided.

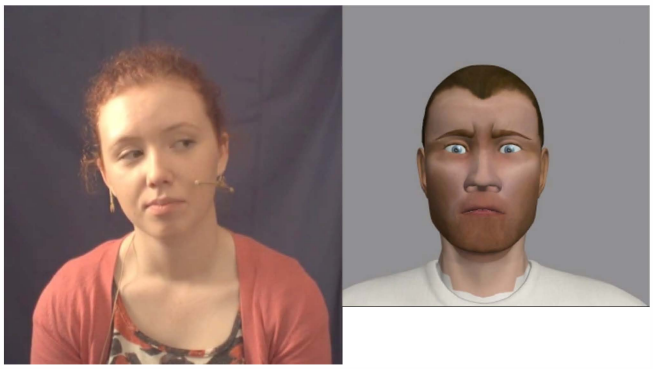

Fig. 4. Sample frames from the AVEC 2011 dataset (the avatar's video was not part of our experiments).

The audio and video data were preprocessed as in [15]. For the video data, each video sequence was pre-processed using the Omron OKAO Vision software library [21] to extract the following facial features: horizontal eye gaze direction (degrees), vertical eye gaze direction (degrees), smile intensity (from 0-100), and head tilt (degrees). We believe that our model will better extract the non-linear relationship between these often geometrical features than the previously existing models.

Similarly, the audio data were also preprocessed using correlation-based feature selection (CFS) [22] to obtain a smaller subset of features more relevant to the task. Since sampling frequencies were not the same for the audio and video signals, the video features extracted were averaged over the course of each word to produce sequences of the same length. Other alternatives to combining these signals could have been evaluated, but this is not the topic of this work. Furthermore, every model evaluated used the same input features and thus the same fusion technique, assuring a fair comparison.

A. Models

The LDCNF model is compared against four other models: Conditional Random Field (CRF), Support Vector Machine (SVM [23]), Latent-Dynamic Conditional Random Field (LDCRF), and Conditional Neural Field (CNF). Comparison of the different models used for evaluation is done in Fig. 3.

Conditional Random Field: As a baseline, the performance of a linear-chain CRF is compared against the other methods. Long-range dependencies were used for the input features – the model was tested for window sizes in $\{0, 1, 2\}$. A window of size 1 implies that each label is predicted by looking at, in addition to the current data sample, both the one before and after (for a window of size 2, two before and two after, etc.). Different regularization parameter values were considered, in the range 10^k, with $k = \{-2, -1, \ldots, 2\}$.

Support Vector Machine: The second baseline is a the multi-class SVM trained on independent frames with a Radial Basis Function (RBF) kernel. During training and validation, two parameters were validated: C, the penalty for classification errors, and γ, a parameter of the RBF function, both with values 10^k, $k = \{-2, -1, \ldots, 2\}$.

Latent-Dynamic Conditional Random Field: Naturally, the LDCRF's performance is evaluated on the given problem. Similarly to the CRF, long-range dependencies are considered (window size $\in \{0, 1, 2\}$) and different regularization parameter values are tested $\sigma = 10^k$, $k = \{-2, -1, \ldots, 2\}$. Various number of hidden states are also considered, $hs \in \{2, 3, 4\}$.

Conditional Neural Field: We compare against the simpler CNF method, to assert whether or not LDCNF offers a better performance for its two hidden layers. In this case, the parameters to test for are window sizes (in $\{0, 1, 2\}$), regularization parameter values ($\sigma = 10^k, k = \{-2, -1, \ldots, 2\}$), and the number of gates to use (in $\{3, 4, 5, 10\}$).

Latent-dynamic Conditional Neural Fields: Performance was computed was computed for the LDCNF model for different values of window size ($\in \{0, 1, 2\}$), regularization parameter ($10^k, k = \{-2, -1, \ldots, 2\}$), number of hidden states (in $\{2, 3, 4\}$), number of gates (in $\{3, 4, 5, 10\}$), and the additional regularization parameter ($\sigma = 10^k, k = \{-3, -2, \ldots, 0, 1\}$).

B. Methodology

For all methods, hold-out testing and validation sets were used. Training was performed on a set of 31 sequences, with validation and testing on separate datasets of 16 sequences each. In the terminology used by the AVEC dataset maintainers, the *training* dataset was kept intact and used solely for training, while the *development* dataset was split in half, one half becoming the validation dataset, the other half becoming the testing dataset.

For each method, on each dataset, the optimal parameters were selected based on F1 performance (or F-measure) on the validation dataset. The F1 score is given by :

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{16}$$

For models where optimization is non convex (LDCRF, CNF, LDCNF) three random starts were issued for each parameter set, and the best one (based on the performance on the validation set) was used for selection of the best parameters.

TABLE I
ERROR MEASURES FOR THE MODELS OF INTEREST.

	AUC	EER	F1
CRF	60.01	58.79	56.68
SVM	70.71	65.84	65.21
LDCRF	75.81	69.48	67.95
CNF	88.71	80.54	79.90
LDCNF	**91.63**	**83.99**	**82.78**

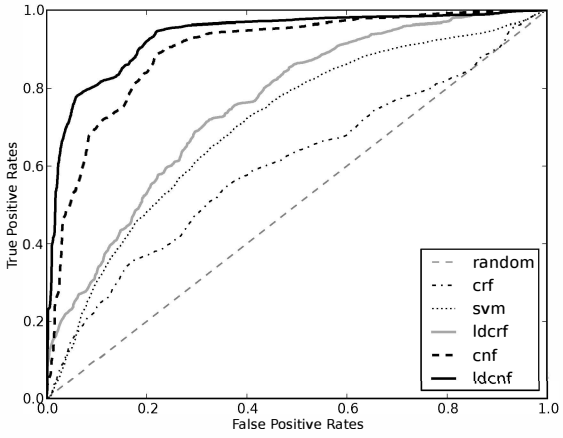

Fig. 5. ROC curves for the best models found during parameter search.

VI. RESULTS AND DISCUSSION

In this section, we present the area under the ROC curve, equal error rate and F1 accuracy for each method combined in Table I. These results show that modeling the label sub-structure using hidden states always improves the performance. This is shown both by comparing the LDCRF with the CRF model and by comparing the LDCNF model with the CNF model. The results also confirm that modeling the non-linearity between input features and labels using neural network improves performance. This is shown by both the CNF vs. CRF comparison and the LDCNF and LDCRF comparison. By integrating both latent layers, our LDCNF model outperforms all previous approaches. Fig. 5 also shows the ROC curves for the different trained models.

To better understand the possible impact of the regularization factor added in Equation 9, we study the performance of the different LDCNF models trained during our parameter search with regards to the α parameter value. Performances are drawn in Fig. 6. From this graph, the best α value would seem to be located around 0.1, and it was observed in our tests that this value provided the better performance. This $alpha$ parameter forced distinct hidden-states to be modeled and seemed benefical up to a point, where the regularization became too strong and started hindering the optimization process.

The fact that our regularization seems to improve performance is an interesting result, and further research should investigate the impact of this regularization on more problems -

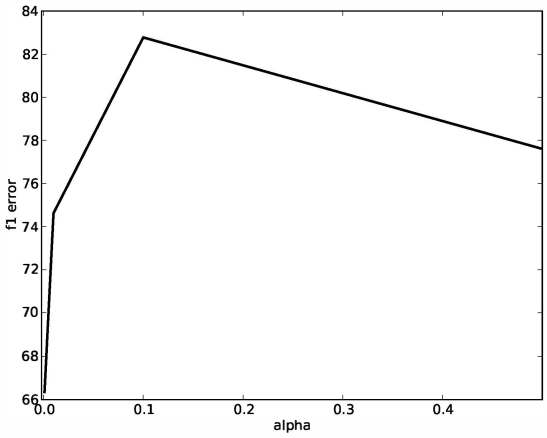

Fig. 6. Testing performance for the LDCNF model on AVEC dataset with varying values of α parameter.

as well as investigating the use of different regularization terms.

VII. Conclusion

In this paper, we presented a model for the labeling of unsegmented data sequences, applied to audiovisual emotion recognition. This model uses two hidden layers, the first to extract a better representation of the input data, and a second to model temporal sub-structures in the sequences at hand. A new regularization term is proposed to constrain the training of the hidden states, encouraging them to be different. Our experiments have shown that this model improves performance over previous methods, and that the introduced regularization term is beneficial for training.

Further work should study what other techniques can be used to make the training of this type of model more straightforward, including different regularization terms and layer-wise training. We also plan to evaluate our model on other video activity recognition that require both the modeling of temporal sub-structure and the extraction of non-linear relationships between the input data and output labels.

References

[1] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, 2001, pp. 282–289.

[2] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2005, pp. 1808–1815.

[3] Y. Song, D. Demirdjian, and R. Davis, "Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2011, pp. 500–506.

[4] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2011 - the first international audio/visual emotion challenge," in *Proceedings of the First International Audio/Visual Emotion Challenge and Workshop (AVEC)*, 2011.

[5] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–53, 2007.

[6] L. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.

[7] J. Peng, L. Bo, and J. Xu, "Conditional Neural Fields," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1419–1427.

[8] L. van der Maaten, M. Welling, and L. Saul, "Hidden-Unit Conditional Random Fields," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, no. 10, 2011, pp. 1848–1852.

[9] A. Shyr and R. Urtasun, "Sufficient dimension reduction for visual sequence classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[10] J. Tao and T. Tan, "Affective computing: A review," in *Affective Computing and Intelligent Interaction*. Springer Berlin Heidelberg, 2005, pp. 981–995.

[11] Z. Zeng and M. Pantic, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[12] M. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 3695–3699.

[13] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies," in *Ninth Annual Conference of the International Speech Communication Association (Interspeech)*, 2008, pp. 597–600.

[14] F. Eyben, M. Wollmer, M. Valstar, H. Gunes, B. Schuller, and M. Pantic, "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect," in *IEEE Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, 2011, pp. 322–329.

[15] G. Ramirez, T. Baltrušaitis, and L. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," *Fourth International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 396–406, 2011.

[16] S. Jain, C. Hu, and J. K. Aggarwal, "Facial Expression Recognition with Temporal Modeling of Shapes The University of Texas at Austin," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 1642–1649.

[17] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2634–2641.

[18] D. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical programming*, 1989.

[19] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*, 2003, pp. 134–141.

[20] G. McKeown and M. Valstar, "The SEMAINE corpus of emotionally coloured character interactions," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2010, pp. 1079–1084.

[21] "OKAO Vision - Omron Tech," http://www.omron.com.

[22] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[23] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1999.