Modeling Latent Discriminative Dynamic of Multi-Dimensional Affective Signals

Geovany A. Ramirez[1], Tadas Baltrušaitis[2], and Louis-Philippe Morency[3]

[1] Computer Science Department, University of Texas at El Paso, USA
garamirez@miners.utep.edu
[2] Computer Laboratory, University of Cambridge, United Kingdom
tadas.baltrusaitis@cl.cam.ac.uk
[3] Institute for Creative Technologies, University of Southern California, USA
morency@ict.usc.edu

Abstract. During face-to-face communication, people continuously exchange para-linguistic information such as their emotional state through facial expressions, posture shifts, gaze patterns and prosody. These affective signals are subtle and complex. In this paper, we propose to explicitly model the interaction between the high level perceptual features using Latent-Dynamic Conditional Random Fields. This approach has the advantage of explicitly learning the sub-structure of the affective signals as well as the extrinsic dynamic between emotional labels. We evaluate our approach on the Audio-Visual Emotion Challenge (AVEC 2011) dataset. By using visual features easily computable using off-the-shelf sensing software (vertical and horizontal eye gaze, head tilt and smile intensity), we show that our approach based on LDCRF model outperforms previously published baselines for all four affective dimensions. By integrating audio features, our approach also outperforms the audio-visual baseline.

Keywords: audio-visual emotion recognition, multi-modal fusion, latent variable models, conditional random fields

1 Introduction

Automated recognition and analysis of human emotions is an important part of the development of affect sensitive AI systems [18]. Humans display affective behavior that is multi-modal, subtle and complex. People are adept at expressing themselves and interpreting others through the use of such non-verbal cues as vocal prosody, facial expressions, eye gaze, various hand gestures, head motion and posture. All of these modalities contain important affective information that can be used to automatically infer the emotional state of a person [23, 7].

Majority of work in automated emotion recognition so far [23] has focused on analysis of the six discrete basic emotions [4] (happiness, sadness, surprise, fear, anger and disgust), even though in everyday interactions people exhibit non-basic and recognisable mental/affective states such as interest, boredom, confusion etc. [20].

Furthermore, because a single label (or multiple discrete labels from a small set) might not describe the complexity of an affective state well, there has been a move to analyse emotional videos/audio along a set of small number of latent dimensions, providing a continuous rather than a categorical view of emotions. Examples of such affective dimensions are power (sense of control), valence (pleasant vs. unpleasant), activation (relaxed vs. aroused), and expectancy (anticipation). Fontaine *et al.* [6] argue that these four dimensions account for most of the distinctions between everyday emotion categories, and hence form a good set to analyse.

Automatic affect sensing and recognition researches have started exploring this venue as well [7]. The problem of dimensional affect recognition is often posed as a binary classification problem [7] (active vs. passive, positive vs. negative etc.) or even as a four-class one (classification into quadrants of a 2D space) rather than a regression one, although there are some exceptions (see Section 2 for more details). In our work we represent the problem as a separate binary classification one along each of the four dimensions.

In addition, most of the work so far has concentrated on analysing different modalities in isolation rather than looking for ways to fuse them [23, 7]. This is partly due to the limited availability of suitably labeled multi-modal datasets and the difficulty of fusion itself, as the optimal level at which the features should be fused is still an open research question [23, 7].

We present a Latent-Dynamic Conditional Random Field [13] (LDCRF) based model to infer the dimensional emotional labels from multiple high level visual cues and a set of auditory features. This approach has the advantage of explicitly learning the sub-structure of the affective signals as well as the extrinsic dynamic between emotional labels. The dimensions analysed in our work are power, valence, expectancy, and activation. Our model is evaluated on the First International Audio/Visual Emotion Challenge (AVEC 2011) dataset. A complete description of the challenge and the dataset can be found in Shuller *et al.* [21]. For the challenge the originally continuous dimensions were redefined as binary ones based on whether they were above or below average, this reduced a regression problem into a classification one.

We evaluate our method on all of the three challenge datasets: video, audio and audio-visual. This allows us to examine the suitability of our approach for analysing audio, visual and audio-visual data. We see an improvement in performance over all of the selected baselines (Support Vector Machines, Conditional Random Fields and Decision Trees) when evaluating our approach on the development set. Furthermore, when evaluated on the test set our approach improves the baseline results for video and audio-visual data.

2 Previous Work

As this paper concentrates on recognition of emotion in dimensional space we present the previous work on this specific task. For recent surveys of dimensional and categorical affect recognition see Zeng *et al.* [23], and Gunes and Pantic [7].

Of special relevance to our work is the work done by Wöllmer *et al.* [22] that uses Conditional Random Fields (CRF) for discrete emotion recognition by quantising the continuous labels for valance and arousal 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.

Nicolaou *et al.* [14] present experiments for classification of spontaneous affect based on Audio-Visual features using coupled Hidden Markov Models that allow them to model temporal correlations between different cues and modalities. They also show the benefits of using the likelihoods produced from separate (C)HMMs as input to another classifier, rather than picking the label with a maximum likelihood for audio-visual classification of affective data. Interestingly, their experiments show that visual features contribute more in spontaneous affect classification in the valence dimension. As in our work the task is approached as a classification rather than regression one.

Nicolaou *et al.* [15] propose the use of Output-Associative Relevance Vector Machine (OA-RVM) for dimensional and continuous prediction of emotions based on automatically tracked facial feature points. Their work poses the dimensional labeling problem as a regression and not a classification one. Their proposed regression framework exploits the inter-correlation between the valence and arousal dimensions by including in their mode the initial output estimation together with their input features. In addition, OA-RVM regression attempts to capture the temporal dynamics of output by employing a window that covers a set of past and future outputs.

Eyben *et al.* [5] fuse both visual (head motion, facial action units) and audio modalities in order to analyse human affect in valence and expectation dimensions. Their results show improved performance when using high-level event-based features such as smiles, head shakes or laughter rather than low-level signal-based ones such as facial feature points or spectral information when predicting affect from audiovisual data in valence and expectation dimensions.

3 Approach

When approaching the challenging problem of recognizing affective dimensions in un-segmented video and audio sequences, one valid approach is to experiment with an extensive set of visual or audio features, where each feature is a low-level representation of the instantaneous appearance of the face or a low level descriptor of the audio signal. The problem with this approach is that the feature space will end up extremely large (5900 dimensions of visual and 1941 of audio features in the case of Schuller *et al.* [21]). This high dimensionality issue can be partially solved by performing dimensionality reduction or feature selection.

For audio features we employ a standard approach of selecting a subset of features using Correlations-based Feature Selection (CFS) [8]. For visual features however, we propose to take advantage of the existing visual sensing techno-

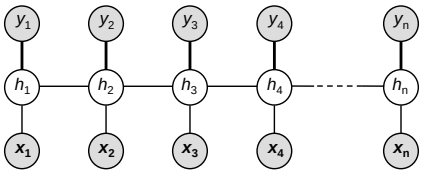


Fig. 1: Graphical representation of the LDCRF model. x_j represents the $j^{\texttt{th}}$ observation (corresponding to the $j^{\texttt{th}}$ observation of the sequence), h_j is a hidden state assigned to x_j , and y_j the class label of x_j (i.e. positive or negative). Gray circles are observed variables.

logy such as Omron OKAO Vision [17] and SHORE[4], to automatically compute higher-level visual features. These commercial and open-source software packages can detect visual features (e.g., eye corners) and recognize high-level communicative signals (e.g., smile intensity). We selected a subset of communicative signals which were shown to be useful when analyzing dyadic interactions [2, 11, 1] and could be estimated robustly: eye gaze, smile and head tilt (see Section 4.2 for more details). By using higher-level visual features, we have the advantage of lower dimensionality, which allows us to learn the interaction between features.

To recognize affective dimensions, we propose to explicitly learn the hidden dynamics between input features (e.g., gaze and smile) using the Latent-Dynamic Conditional Random Field(LDCRF) model [13] (see Figure 1). LDCRF offers several advantages over previous discriminative models. In contrast to Conditional Random Fields(CRFs) [12], LDCRF incorporates hidden state variables which model the sub-structure of gesture sequences. The CRF approach models the transitions between gestures, thus capturing extrinsic dynamics, but lacks the ability to represent internal sub-structure. LDCRF can learn the dynamics between gesture labels and can be directly applied to labeled unsegmented sequences.

As described in Morency *et al.* [13], the task of the LDCRF model is to learn a mapping between a sequence of observations $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ and a sequence of labels $\mathbf{y} = \{y_1, y_2, ..., y_m\}$. Each y_j is a class label for the $j^{\texttt{th}}$ observation in a sequence and is a member of a set $\mathcal{Y}$ of possible class labels, for example, $\mathcal{Y} = \{\texttt{positive-valence}, \texttt{negative-valence}\}$. Each frame observation x_j is represented by a feature vector $\phi(x_j) \in \mathbf{R}^d$, for example, the eye gaze, smile and head tilt at each frame. For each sequence, we also assume a vector of "substructure" variables $\mathbf{h} = \{h_1, h_2, ..., h_m\}$. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model.

Given the above definitions, we define a latent conditional model:

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}} P(\mathbf{y} \mid \mathbf{h}, \mathbf{x}, \theta) P(\mathbf{h} \mid \mathbf{x}, \theta). \tag{1}$$

where θ are the parameters of the model.

[4] `http://www.iis.fraunhofer.de/en/bf/bv/ks/gpe/demo/`

Given a training set consisting of n labeled sequences $(\mathbf{x_i}, \mathbf{y_i})$ for $i = 1...n$, training is done following Lafferty *et al.* [12] using this objective function to learn the parameter θ^*:

$$L(\theta) = \sum_{i=1}^{n} \log P(\mathbf{y}_i \mid \mathbf{x}_i, \theta) - \frac{1}{2\sigma^2}||\theta||^2 \quad (2)$$

The first term in Eq. 2 is the conditional log-likelihood of the training data. The second term is the log of a Gaussian prior with variance σ^2, i.e., $P(\theta) \sim \exp\left(\frac{1}{2\sigma^2}||\theta||^2\right)$.

For a more detailed discussion of LDCRF training and inference see Morency *et al.* [13].

4 Experimental Setup

In this section we first introduce the dataset used for validating and testing our approach. We follow with a discussion of the audio and video features selected for our experiments. We then describe the audio-visual fusion methods used in our experiments. Finally, the training and the validation of the models is described.

4.1 Dataset

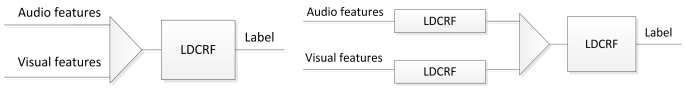
For all our experiments we used the dataset provided by Schuller *et al.* [21]. The dataset consist of 95 video and audio recorded dyadic interaction sessions between human participants and a virtual agent operated by a human. The dataset consists of upper body video segments with per frame and audio and audio-visual segments with per word binary labels along the four affective dimensions (activation, expectation, power and valence).

4.2 Visual Features

As discussed in Section 3, we selected a subset of visual communicative signals which were shown to be useful when analyzing dyadic interactions [2, 11, 1] and could be estimated robustly by an off-the-shelf sensing software. In our experiments, we processed each video sequence with the Omron OKAO Vision software library [17] to automatically 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 reason that the visual features that play an important role in face-to-face communication are potentially good for affective signal recognition. In addition, the use of higher-level visual features of low dimensionality allows us to learn the interaction between features better.

4.3 Audio Features

For our original audio feature set we used the 1941 features provided with the dataset [21]. Each of the features were sampled over a duration of a single word (mean word length is 263ms). As the dimensionality of the feature set is very high we applied Correlation-based Feature Selection (CFS) [8] to select a subset



(a) Early (feature level) fusion **(b)** Late (decision level) fusion

Fig. 2: The two multimodal fusion techniques used in our experiments.

of features relevant for the task. Due to memory limitations of WEKA [9] toolkit a subsample of the audio training set was taken for feature selection (every third word). On the resulting subset a 10-fold cross validation CFS was performed on each of the four emotion labels (valence, activation, expectancy, and power) independently. That is, leaving a 10th of the training set out and running CFS on the remaining data. Features that were chosen in at least 5 of 10 folds were chosen as input features for our model, with the exception of arousal where only the features selected in all of the ten folds were chosen as the 5 out of 10 approach lead to 91 features. This was done in order to keep the dimensionality low, and have a roughly similar number of features across the different affective dimensions. This resulted in 19 features for arousal, 7 for expectancy, 22 for power, and 15 for valence dimensions.

4.4 Audio-Visual Features

In the case of fusion we needed to align the visual with audio features. Audio features were sampled over a longer and varying period of time as opposed to per frame sampling of video features. For computing the visual features at the word level, we used the mean values of all the frames happening during a specific word. This resulted in same length sequences for both video and audio features.

The optimal level of fusion is still an open research question. In our work we explore several approaches fusion (illustrated in Figure 2). The most straightforward one is to concatenate the audio and visual features and train a classifier on them (Figure 2a). An alternative to that is to concatenate the marginal probabilities output from the unimodal models (Figure 2b) and use that as an input to another classifier.

4.5 Baseline Models

In addition to the baseline provided by Shuller *et al.* [21], we decided to evaluate our approach (LDCRF) against Conditional Random Fields (CRF), which were already used in an affective dimension classification task by Wöllmer *et al.* [22]. We also compare our approach to decision trees due to their simplicity and speed of training, and also to provide us with an additional non-temporal model as a baseline.

SVM The baseline proposed by Shuller et al. [21] uses Support Vector Machine (SVM) classification without feature selection. For the audio data a linear kernel SVM was used, while for the video a radial basis function kernel SVM.

Both were trained on the training dataset. For the audiovisual data, they used late (decision level) fusion using a linear SVM trained on the development set.

Decision Trees We used the Java implementation of C4.5 algorithm for decision trees as our first baseline model [19]. The decision tree is created based on the information entropy of each feature in a given training set.

Conditional Random Field For our third baseline we trained a single Conditional Random Field (CRF) for each affective dimension using the visual input features [12]. The CRF model has a similar structure as the LDCRF model but without the hidden variables. No latent dynamic is explicitly learned with the CRF model.

4.6 Methodology

For all the experiments we use the data provided by Schuller *et al.* [21]. The data is divided into 3 subsets: training, development and testing. The training set consists of 31 sessions, while the development set consists of 32 sessions that were used for validation of the model parameters. The test set consists of 11 video only sequences, 11 audio only sequences, and 10 audio-visual sequences. All of the 32 test sequences did not have any publicly available labels. The same validation and testing methodology was applied to audio, video and audio-visual data. In the case of late fusion the training was again performed on the training dataset and validated on development one (same technique used for the unimodal models). In each case a separate binary classifier is trained for each of the dimensions rather than one giving multiple-label output. We automatically validated the following model parameters: for CRF and LDCRF the L2-norm regularization term was validated with values $10^k, k = -2..3$, for LDCRF we validated the number of hidden states (2-4), no validation was performed for decision trees. We used weighted accuracy as the measure of performance for all of our experiments.

For the decision tree experiments we used the WEKA toolkit [9]. For training CRF and LDCRF we used the freely available hCRF library [10].

5 Results and Discussion

5.1 Visual data

The goal of our experiments on visual data were two fold: evaluating the selected visual features (gaze, smile and head tilt) and comparing our approach based on LDCRF with other baseline models. First, we performed a comparison of our selected set of features (see Section 4.2) with the Local Binary Patterns [16] used in Schuller *et al.* [21]. For fair comparison, we trained a Support Vector Machine (SVM) with a radial basis function (RBF) kernel as was performed by Schuller *et al.* [21], who used a 5900 dimensional visual feature vector, while our feature vector was only 4 dimensional (horizontal and vertical gaze, smile and head tilt). The performance of using our features can be seen (Table 1) to be similar to those used by Shuller *et al.* [21], showing that our selected features are at at least as good as theirs.

Table 1: Visual feature evaluation: comparison between the use of Local Binary Patterns features [21] and our set of features described in Section 4.2 on the development set.

Weighted accuracy (%)	**Activation**	**Expectancy**	**Power**	**Valence**
SVM + LBP [21]	60.2	58.3	56.0	63.6
SVM + Our features	**58.7**	**60.3**	**54.0**	**63.6**

Table 2: Classification results of our approach (LDCRF) and the baselines on the development dataset for audio and video modalities. A stands for activation, E for expectancy, P for power and V for valence.

Weighted accuracy	Audio					Video				
(%)	A	E	P	V	Average	A	E	P	V	Average
Baseline [21]	63.7	63.2	65.6	58.1	62.7	60.2	58.3	56.0	63.6	59.5
Decision trees	60.8	65.9	63.1	**64.3**	63.5	62.1	55.4	49.3	64.1	57.7
CRF	62.9	67.3	67.0	44.6	60.4	72.3	53.8	46.2	69.5	60.5
LDCRF	**74.9**	**68.4**	**67.0**	63.7	**68.5**	**74.5**	**60.0**	**60.3**	**72.9**	**66.9**

We compare our approach with the three selected baselines (described in Section 4.5) on the development set. It can be seen from Table 2, video sub-challenge, that our LDCRF approach outperforms all of them in all of the affection dimensions. We also evaluate our LDCRF model on the test set, it can be seen to outperform the baseline [21] in all of the affective dimensions (Table 4, video sub-challenge). We only compare the LDCRF method on the test set due to a limited number of five attempts of result submissions per sub-challenge.

5.2 Audio Data

Similarly to the video data experiments we first evaluate the approach on audio data against all three of the baselines on the development set (Table 2, audio sub-challenge), and only the LDCRF model on the test set (Table 4 audio sub-challenge). On the development set our model can be seen outperforming other approaches. On the test set, LDCRF performance is similar to the SVM baseline. The low performance of both SVM and LDCRF approaches (e.g. the best performance on the Power labels is 28%) on this test set suggests a significant difference in the data distribution of the audio-only sub-challenge test set.

5.3 Audio-Visual Data

For audio-visual fusion two experiments were performed, comparing the fusion methods and evaluating them on the test dataset. From Table 3 we can see that the late fusion using LDCRF as a model to fuse the outputs of uni-modal classifiers performed best in all of the affective dimensions, so only this approach was evaluated on the test set.

Table 3: Fusion methods using LDCRF classifiers on the development set.

Weighted accuracy (%)	**Activation**	**Expectancy**	**Power**	**Valence**	**avg.**
Early Fusion (LDCRF)	79.3	63.4	66.9	62.8	68.1
Late Fusion (LDCRF)	**81.7**	**73.1**	**73.3**	**73.5**	**75.4**
Late Fusion (SVM)	75.4	69.4	65.3	72.1	70.5

Table 4: Official results on the test set.

Weighted accuracy (%)	**Activation**	**Expectancy**	**Power**	**Valence**	**avg.**
Video sub-challenge					
Baseline [21]	42.2	53.6	36.4	52.5	46.2
LDCRF	**65.5**	**61.7**	**47.1**	**69.8**	**61.0**
Audio sub-challenge					
Baseline [21]	55.0	**52.9**	**28.0**	44.3	**45.1**
LDCRF	**55.8**	50.1	19.8	**46.5**	43.0
Audiovisual sub-challenge					
Baseline [21]	**67.2**	36.3	62.2	**66.0**	57.9
LDCRF	65.6	**53.4**	**62.9**	59.5	**60.3**

We can see in Table 4 the results of our approach on the audio-visual test set. Our LDCRF based approach gives results that are always better than 50% (chance level) while the SVM baseline approach performs below chance level for expectancy. Our LDCRF based approach outperforms the SVM-based baseline on the average over all 4 emotion labels.

6 Conclusion

In this paper, we proposed an approach that models the interaction between the high level perceptual features using Latent-Dynamic Conditional Random Fields. We evaluated our approach on the Audio-Visual Emotion Challenge (AVEC 2011) dataset. By using visual features easily computable using off-the-shelf software, we showed that our approach based on LDCRF model outperforms the previously published baseline for all four affective dimensions on both the development and test datasets. Integrating this with audio data we are able to improve performance of the baseline on audio-visual data on average.

LDCRF model seems to be suitable for late feature fusion outperforming the SVM model for fusion. Looking at the results from video and audio sub-challenges, they did not seem to generalise very well on the audio-visual sub-challenge data (both in the case of baseline and our approach), this might have been due to the differences between the test subsets. This points to the need of future work looking into using semi-supervised domain adaptation techniques (such as proposed by Blitzer *et al.* [3]) to learn the distribution of features in unseen data allowing the methods to generalise better.

References

1. Argyle, M., Dean, J.: Eye-contact, distance and affiliation. Sociometry 28, 233–304 (1965)
2. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology 79(6), 941–952 (2000)
3. Blitzer, J., McDonald, R., Pereira, F.: Domain Adaptation with Structural Correspondence Learning. In: EMNLP. pp. 120–128 (2006)
4. Ekman, P.: An argument for basic emotions. Cognition & Emotion 6(3), 169–200 (1992)
5. Eyben, F., Wollmer, M., Valstar, M., Gunes, H., Schuller, B., Pantic, M.: String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In: IEEE FG'11) (2011)
6. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.: The world of emotion is not two-dimensional. Psychological Science 18, 1050–1057 (2007)
7. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int'l Journal of Synthetic Emotion 1(1), 68–99 (2010)
8. Hall, M.: Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, University of Waikato (1999)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11, 10–18 (2009)
10. HCRF: library for crf and ldcrf `http://sourceforge.net/projects/hcrf/`
11. Krämer N. C.: Human behavior in military contexts, chap. Nonverbal Communication, pp. 150 – 188. Washington: The National Academies Press (2008)
12. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labelling sequence data. In: ICML (2001)
13. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: CVPR '07 (2007)
14. Nicolaou, M., Gunes, H., Pantic, M.: Audio-visual classification and fusion of spontaneous affective data in likelihood space. In: ICPR (2010)
15. Nicolaou, M., Gunes, H., Pantic, M.: Output-associative rvm regression for dimensional and continuous emotion prediction. In: IEEE FG'11 (2011)
16. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. TPAMI 24(7) (2002)
17. OKAO: software `http://www.omron.com/r_d/coretech/vision/okao.html`
18. Pantic, M., Rothkrantz, L.: Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE 91(9), 1370–1390 (2003)
19. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
20. Rozin, P., Cohen, A.B.: High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. Emotion 3(1), 68 – 75 (2003)
21. Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: Avec 2011 the first international audio/visual emotion challenge. In: 1st Int. Audio/Visual Emotion Challenge and Workshop. Springer LNCS (2011)
22. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R.: Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In: INTERSPEECH. ISCA (2008)
23. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. TPAMI 31(1) (2009)