A Multi-layer Hybrid Framework for Dimensional Emotion Classification

Mihalis A. Nicolaou
Department of Computing
Imperial College London, U.K.
mihalis@imperial.ac.uk

Hatice Gunes
EECS, Queen Mary
University of London, U.K.
haticeg@ieee.org

Maja Pantic
Imperial College London, U.K.
U. of Twente, The Netherlands
m.pantic@imperial.ac.uk

ABSTRACT

This paper investigates dimensional emotion prediction and classification from naturalistic facial expressions. Similarly to many pattern recognition problems, dimensional emotion classification requires generating multi-dimensional outputs. To date, classification for valence and arousal dimensions has been done separately, assuming that they are independent. However, various psychological findings suggest that these dimensions are correlated. We therefore propose a novel, *multi-layer hybrid framework* for emotion classification that is able to *model inter-dimensional correlations*. Firstly, we derive a *novel geometric feature set* based on the (a)symmetric spatio-temporal characteristics of facial expressions. Subsequently, we use the proposed feature set to train a multi-layer hybrid framework composed of a *temporal regression layer* for predicting emotion dimensions, a *graphical model layer* for modeling valence-arousal correlations, and a final *classification and fusion layer* exploiting informative statistics extracted from the lower layers. This framework (i) introduces the *Auto-Regressive Coupled HMM* (ACHMM), a graphical model specifically tailored to accommodate not only inter-dimensional correlations but also to exploit the internal dynamics of the actual observations, and (ii) replaces the commonly used Maximum Likelihood principle with a more robust final classification and fusion layer. Subject-independent experimental validation, performed on a naturalistic set of facial expressions, demonstrates the effectiveness of the derived feature set, and the robustness and flexibility of the proposed framework.

Categories and Subject Descriptors

H.1.2 [**User / Machine systems**]: [Human information processing]; I.2.10 [**Vision and scene understanding**]: [Video analysis]; I.5.2 [**Design Methodology**]: [Feature evaluation and selection, Classifier design and evaluation]

* Area Chair: Massimo Zancanaro

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

General Terms

Human Factors, Feature evaluation and selection, Classifier design and evaluation, Experimentation

1. INTRODUCTION

Traditionally, research in the field of automatic emotion recognition has focused on recognizing discrete, basic emotional states (e.g. happiness, sadness) from posed data acquired in laboratory settings [4]. However, these models are deemed unrealistic as they are unable to capture the non-basic and subtle affective states exhibited by humans in everyday interactions. In order to accommodate such subtle expressions, researchers have started adopting a dimensional description of human emotion where an emotional state is characterized in terms of a number of latent dimensions [10]. Two dimensions are deemed sufficient for capturing most of the affective variability: valence and arousal (V-A), signifying respectively, how negative/positive and active / inactive an emotional state is.

Due to the aforementioned reasons, automatic, dimensional and continuous emotion recognition has increasingly attracted the interest of the affective computing researchers in recent years. Wollmer et al. [13], quantize the V-A dimensions into 4 or 7 levels and use Conditional Random Fields (CRFs) for classification from audio cues. Other works discriminate the emotions into more coarse classes such as positive vs. negative [9], by combining audio-visual cues via Coupled Hidden Markov Models (CHMMs) and likelihood space fusion, while the work of [14] utilizes a dynamic Bayesian network combined with Long-Short Term Memory (LSTM) Neural Nets, performing regression and quantizing the results into four quadrants.

Various findings in psychology (i) postulate the existence of inter-dimensional correlations between the valence and arousal dimensions [8, 7], and (ii) suggest that there exist asymmetrical variations on the human face (left vs. right, lower vs. upper hemiface) depending on whether the subject is expressing a positive or a negative emotion [2, 11]. Although such features have been used for other human behavior understanding problems they have not been explored for dimensional emotion prediction. The work presented in this paper takes into account these findings, and differs from the previous works on automatic dimensional emotion prediction by proposing the following contributions.

(A)symmetric features. We derive a set of geometric spatio-temporal meta-features from facial feature point trackings. We show how the derived feature set enables better prediction accuracy compared to the initial set of raw fea-

tures (i.e., facial feature point locations).

Auto-Regressive Coupled Hidden Markov Models (ACHMM). We introduce the ACHMM, a novel HMM variant specifically tailored for capturing inter-dimensional patterns and observation dynamics. To the best of our knowledge, this paper pioneers the use of ACHMM.

A multi-layer hybrid framework. We propose a novel multi-layer hybrid classification framework composed of three distinct layers: (1) A *regression layer* generating the continuous A-V prediction (using LSTM), (2) a *graphical model layer* trained on the predicted emotion dimensions to capture the correlations between the continuous emotion descriptions (using ACHMM), and (3) a final *discriminative classification and fusion layer* (using SVM) for incorporating informative statistics extracted from both the regression and the graphical model layer, replacing the commonly applied (but less effective) Maximum Likelihood principle.

More specifically, the proposed framework aims to capture specific, intra-dimensional statistics via the regression layer, and subtle, emerging inter-dimensional patterns via the graphical model layer. Outputs from both layers are combined via a final discriminative classification layer that estimates a set of statistics over the previous two layers and uses this information for *classifier fusion*. To date, no such work has been attempted for dimensional emotion classification. Subject-independent experimental validation, performed on a naturalistic set of facial expressions, demonstrates the effectiveness of the derived feature set, and the robustness and flexibility of the proposed multi-layer hybrid framework.

2. DATA SET

We used the Sensitive Artificial Listener (SAL) Database [3] for this work. It contains naturalistic audio-visual conversational data taking place between a participant and a human operated avatar. Each avatar has a different personality (happy, gloomy, angry or pragmatic). The recordings were made in controlled laboratory settings with one camera, microphones, uniform background, and constant lightning conditions. Only data from 4 subjects have been continuously annotated by 3-4 coders along the V-A dimensional (emotion) space. Representative frames together with their facial point trackings are shown in Fig. 1. In total, we used 61 positive and 73 negative episodes ($\approx$ 30,000 frames) capturing transitions to an emotional state and back (e.g., going from nonpositive to positive and back to non-positive).

3. FEATURE EXTRACTION & SELECTION

In this section, we explain how we process and select the facial features used for training the first-layer regressors.

Similarly to [12], we track 20 facial feature points which represent the facial expression of the subject in any given frame (see Fig. 1). This is referred to as the *original feature set*. We register each set of points in a given frame to the corresponding coordinate system centered at the fixed point of the face (the average of the inner eye points and the tip of the nose). We thus end up with a simple translation applied to every point in every frame (also using the fixed point itself as a feature).

Motivated by psychological findings [1, 2, 11], we extract various symmetrical features, in an attempt to capture facial asymmetry manifestations. These features include distances

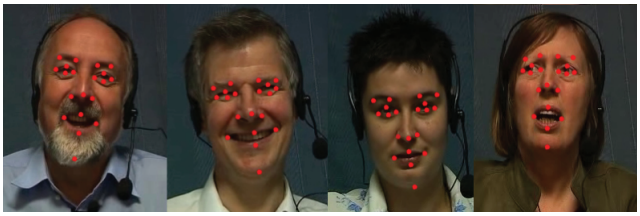

Figure 1: Examples from the SAL data set along with the extracted 20 points, used as (original) features for facial expression recognition.

between points on the left vs. right hemiface, and the upper vs. lower hemiface (e.g., the distance between left & right eyebrow, between upper & lower eyelid, etc.). For each of these symmetric pairs of points (p_l, p_r), we extract the angle, the coordinate distance $\langle p_l - p_r \rangle$, and the Euclidean distance $(||p_l - p_r||)$. Subsequently, we derive a new set of features to capture the temporal aspect of a given expression. For every point $p =< x_p, y_p >$ we extract the velocity in x and y directions: $< \frac{dx_p}{dt}, \frac{dy_p}{dt} >$. We also fit a set of overlapping windows of length w to an n-degree polynomial $y^w(x^w) = a_n x^n + \cdots + a_0$ (t represents the current frame, and the window is defined as $[t-w, t+w]$). By doing so, we derive the $n+1$ coefficients ($[a_n \ldots a_0]$). In summary, the *full feature set* (fs) consists of facial feature point coordinates registered to the fixed point ($< x, y >_{rfp}$), a vector defined by subtracting the two symmetric points $< x, y >_s$, symmetric angles (π), symmetric distances ($||.||$), velocities ($\frac{dp}{dt}$), and polynomial fitting coefficients ($a_n t^n$).

In order to select the appropriate features for the problem at hand, we apply Correlation Based Feature Selection (CFS) [5] on the full feature set. The CFS ranking results show that the top ranked features correspond to velocity, point (a)symmetry and polynomial coefficients, signifying the importance of the temporal and (a)symmetric characteristics of facial expressions. We conducted further experiments with LSTM Neural Nets (LSTM-NN) to verify this finding. The obtained results showed that using *the derived feature set* reduced the Mean Squared Error (MSE) both for valence and arousal. Our experiments also showed that including the polynomial coefficients caused the accuracy to decrease.

4. PREDICTION & CLASSIFICATION

The final feature set (omitting polynomial features) obtained is used to train a multi-layer hybrid prediction and classification framework. This 3-layer hybrid framework incorporates a prediction layer for each emotion dimension, a graphical model layer for capturing inter-dimensional emotion correlations, and a final discriminative classification layer incorporating statistics extracted from both layers. In the following sections we firstly describe the employed models, and subsequently explain the structure of the proposed hybrid framework.

4.1 Long-Short Term Memory Neural Nets

LSTM Neural Nets (LSTM-NN) [6] are a form of Recurrent Neural Networks (RNN), which in contrast to traditional RNNs enable the learning of temporal information longer than a few time steps. In LSTM-NNs a typical node

is replaced with a memory cell. The cell maintains the given state of the network, which is considered to be representative of the *previous* (and the *future*, in the bidirectional case) sequence inputs. A set of gates provide 'read, write, reset' operations on the cell state.

4.2 Auto-Regressive Coupled HMM

By merging two common HMM variants, the Coupled HMM and the Auto-Regressive HMM, we propose the Auto-Regressive Coupled HMMs (ACHMM) for the problem of capturing inter-dimensional emotion correlations and structure. The ACHMM is a Dynamic Bayesian Network modeling observation dependencies (not only on the abstract state level, but also conditioned on the actual observations) and capturing the inter-stream structure. CHMMs are structured specifically to model interactions between multiple processes. Assuming that we have two streams of observations, at each time t, we have two hidden nodes and two nodes modeling each observation stream. The state of each hidden node at time t depends on the hidden states of both hidden nodes at $t-1$. By augmenting CHMM with auto-regression, we relax the assumption that the observation depends only on the current state of the model. The distribution of the observation at each time t is now conditioned both on the current hidden state as well as the previous observation. An ACHMM with two streams of observations is illustrated in Fig. 2(a) (see the HMM nodes).

4.3 The Multi-layer Hybrid Framework

The multi-layer hybrid framework we propose is illustrated in Fig. 2. The framework has three distinct layers: (1) a *regression layer* (LSTM-NN) which generates the continuous prediction for each emotion dimension, (2) a *graphical model layer* (ACHMM) that models the inter-dimensional structure, and (3) finally, a *discriminative classification layer* (SVM) incorporating statistics from the lower layers. More specifically, the regression layer is expected to capture specific, intra-dimensional statistics which would serve as intermediate features for more accurate classification of an emotion dimension. The graphical model layer on the other hand, is able to capture subtle, emerging inter-dimensional patterns which seamlessly contribute to the classification of a given sequence.

Let us consider the **LSTM-ACHMM$_{\mathbf{ML}}$** model. Firstly, the LSTM-NNs are trained using the optimal feature set obtained in Section 3. Two separate LSTM-NNs are trained as regressors, one for each emotion dimension. Let $\mathcal{D}$={Valence, Arousal} be the set of emotion dimensions. The continuous prediction generated by each LSTM-NNs at time t represents the observation modeled by each component of the ACHMM at time t. Due to the structure of the ACHMM, for $t=[2,N]$, the observation O_{dt} generated for each emotion dimension $d \in \mathcal{D}$ depends directly on the previous observation for this dimension O_{dt-1}, and the state of the hidden node H_{dt}. Furthermore, the hidden state H_{dt} depends on the previous hidden states of both dimensions (H_{At-1} and H_{Vt-1}) due to the coupled nature of the model. Classification is obtained using Maximum Likelihood (ML) principle.

In order to combine the distinct qualities of the aforementioned models, we add a final *classification and fusion layer* to the framework. This layer exploits a set of statistics from the two lower layers, and uses these features to the aim of classifier fusion using SVMs. Thus, it not only *replaces* the ML-based classification with the discriminative classification of SVMs, but also provides a more robust learning of inter-dimensional patterns and structure via classifier fusion.

Let us now consider the 3-layer hybrid model of **LSTM-ACHMM$_{\mathbf{SVM}}$**. Firstly, from the LSTM-NN prediction, for each dimension $d \in \mathcal{D}$ and sequence s, we extract the following feature vector $f_{\text{LSTM},d}(s)$:

$$\langle \bar{pos} f_d(s), \bar{neg} f_d(s), \bar{sum}_d(s) \rangle$$

where $\bar{pos} f_d(s)$ and $\bar{neg} f_d(s)$ correspond to the percentage of frames with positive/negative output for dimension d, and $\bar{sum}_d(s)$ is the average value of the (sequence) output for this dimension. From the ACHMMs, we extract a subset of the statistics that characterize the model. Again, for each sequence s and dimension d, we obtain the feature vector $f_{\text{ACHMM},d}(s)$:

$$\langle \hat{ll^+}_d(s), \hat{ll^-}_d(s), \mathbf{\bar{MPE}_d(s)} \rangle$$

Where $\hat{ll^+}_d(s)$ and $\hat{ll^-}_d(s)$ are the normalized (using the sequence length, $\hat{ll}_d(s) = \frac{ll}{|s|}$) class likelihoods generated by the model. $\mathbf{\bar{MPE}_d(s)}$, also known as the most probable explanation, refers to the most probable state that each hidden node is at time t. Out of a total of St_n states, let $|St(h_i, St_j)|$ be the number of time steps that the hidden node h_i is at state St_j. Then for each $\bar{St}_j$, a new feature (representing the percentage of time that the hidden node is at every state) is generated as $\bar{St}_j = \frac{|St(h_i,St_j)|}{|s|}$. The feature vector fed into the SVM classifier is described as:

$$\langle f_{\text{LSTM,Val}}(s), f_{\text{LSTM,Ar}}(s), f_{\text{ACHMM,Val}}(s), f_{\text{ACHMM,Ar}}(s) \rangle$$

Notice that statistics extracted for both emotion dimensions are fed into the classifier, thus enabling more robust learning of inter-dimensional patterns and structure.

5. EXPERIMENTAL SETUP & RESULTS

For our experiments, we use the bidirectional LSTM-NNs with one hidden layer. The ACHMMs have 3 hidden states for each hidden node, with each observation modeled by a mixture of 2 Gaussians. We use SVMs with an RBF kernel and optimize the parameters via cross-validation on the training set. The proposed multi-layer hybrid framework is evaluated for two classification tasks: (i) V-A hemispheric classification (positive vs. negative for the valence dimension, and active vs. inactive for the arousal dimension), and (ii) V-A quadrant classification (positive/active, negative/active, positive/inactive, and negative/inactive). All experimental evaluation is obtained by performing *leave-one-subject-out cross-validation*. The results obtained are shown in Table 1. The LSTM results refer to the regression results mapped from the LSTM-NNs onto the valence / arousal classes (via majority voting). When we compare these results to the results obtained from the hybrid LSTM-ACHMM$_{ML}$ model (by applying ML over the ACHMM), the LSTM-NN results provide better $F1$ scores for the valence and arousal classes. The quadrant classification results, however, show that the hybrid model improves the classification results (an 8% increase in the F1 score). This finding supports our assumption that modeling inter dimensional correlations helps in capturing (more) subtle class variances. Finally, the proposed multi-layer hybrid framework (LSTM-ACHMM$_{SVM}$) outperforms both its ML coun-

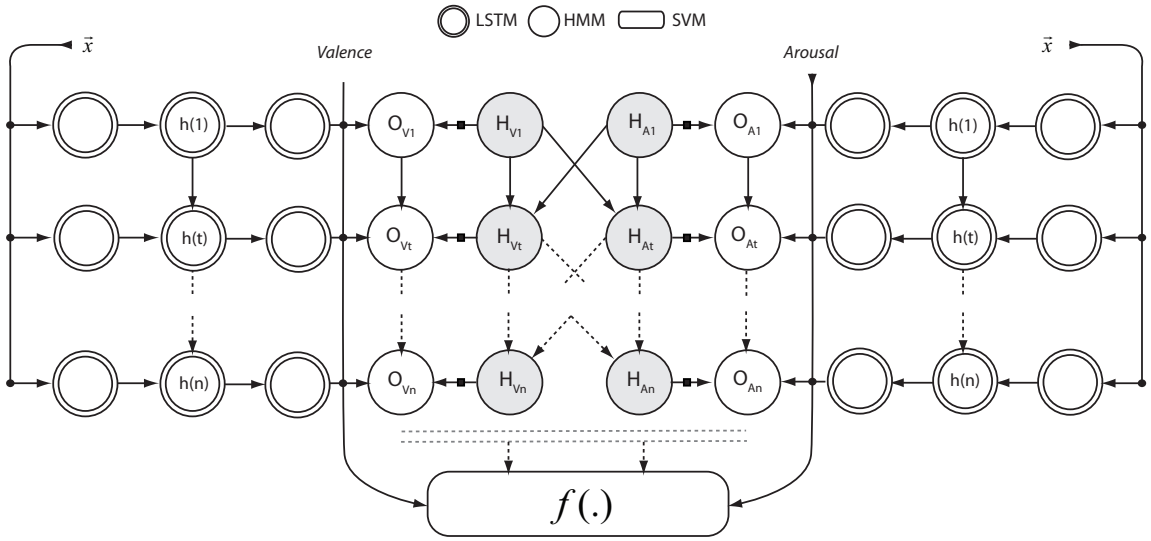

Figure 2: Illustration of the 3-layer hybrid framework employed. The graphical model employed is an ACHMM, where shaded nodes represent hidden nodes and filled squares represent a mixture of Gaussians.

Table 1: Experimental results (ACC: accuracy, PREC: precision, REC: recall) for valence (positive/negative), arousal (active/inactive), and quadrant classification.

Layer(s)	Model	Valence				Arousal				Quadrant			
		ACC	PREC	REC	F1	ACC	PREC	REC	F1	ACC	PREC	REC	F1
layer 1	**LSTM**	80	83	81	80	80	80	75	75	71	60	60	58
layers 1, 2	**LSTM-ACHMM$_{\mathbf{ML}}$**	72	76	70	69	75	67	60	58	67	66	66	66
layers 1, 2, 3	**LSTM-ACHMM$_{\mathbf{SVM}}$**	86	87	86	**86**	84	83	79	**79**	84	90	75	**77**

terpart and the simpler LSTM-based classification in all classification tasks. LSTM-ACHMM $_{SVM}$ achieves an accuracy of 86% and 84% for valence and arousal (hemispheric) classification, respectively, compared to an accuracy of 80% and 71% using LSTM-NNs (for the same classification task). The most significant increase in accuracy is obtained for quadrant classification where the LSTM-ACHMM$_{SVM}$ framework improves both the classification accuracy (from 71% to 84%) and the F1 score (from 58% to 77%).

6. CONCLUSIONS

This paper presented a novel multi-layer hybrid framework that derives (i) symmetric spatio-temporal features, (ii) introduces (and pioneers the use of) ACHMM, and (iii) combines a dynamic regressor (LSTM-NNs), a structured graphical model (ACHMM), and a discriminative classification and fusion scheme (SVM) for subject-independent dimensional emotion prediction and classification. Comparative evaluation, performed on a naturalistic data set of facial expressions, showed that the proposed framework provides an increase in accuracy and robustness in all classification tasks. Specifically, our results indicate that by modeling inter-dimensional covariances we can learn complex and subtle class variances more robustly. Overall, the proposed framework is highly flexible due to its layered nature. It can be easily extended and applied to other multi-dimensional (and/or multi-modal) classification problems.

7. ACKNOWLEDGMENTS

This work has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

8. REFERENCES

[1] J. Borod et al. Facial asymmetry while posing positive and negative emotions. *Neuropsychologia*, 26(5):759 – 764, 1988.
[2] J. Borod et al. Neuropsychological aspects of facial asymmetry during emotional expr. *Neuropsychology Rev.*, 7:41–60, 1997.
[3] E. Douglas-Cowie et al. The HUMAINE database. In *ACII*, pages 488–500, 2007.
[4] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *IJSE*, 1(1):68–99, 2010.
[5] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *ICML*, pages 359–366, CA, USA, 2000.
[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9:1735–1780, November 1997.
[7] R. Lane et al. *Cognitive Neuroscience of Emotion*. Oxford University Press, 2000.
[8] P. A. Lewis et al. Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, 17:742–748, 2007.
[9] M. Nicolaou et al. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *ICPR*, pages 3695–3699, 2010.
[10] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.
[11] B. B. Schiff and B. MacDonald. Facial asymmetries in the spontaneous response to positive and negative emotional arousal. *Neuropsychologia*, 28(8):777 – 785, 1990.
[12] M. Valstar et al. How to distinguish posed from spontaneous smiles using geometric features. In *ICME*, pages 38–45, 2007.
[13] M. Wollmer et al. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Interspeech*, pages 597–600, 2008.
[14] M. Wollmer et al. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE J-STSP*, 4(5):867 –881, 2010.