# Designing Frameworks for Automatic Affect Prediction and Classification in Dimensional Space

Mihalis A. Nicolaou
Imperial College London
United Kingdom
mihalis@imperial.ac.uk

Hatice Gunes
Imperial College London
United Kingdom
h.gunes@imperial.ac.uk

Maja Pantic
Imperial College London, UK
Univ. of Twente, NL
m.pantic@imperial.ac.uk

## Abstract

*This paper focuses on designing frameworks for automatic affect prediction and classification in dimensional space. Similarly to many pattern recognition problems, dimensional affect prediction requires predicting multi-dimensional output vectors (e.g., valence and arousal) given a specific set of input features (e.g., facial expression cues). To date, affect recognition in valence and arousal space has been done separately along each dimension, assuming that they are independent. However, various psychological findings suggest that these dimensions are correlated. In light of this, we focus on modeling inter-dimensional correlations, and propose (i) an Output-Associative Relevance Vector Machine (OA-RVM) regression framework that augments the traditional RVM regression by being able to learn non-linear input and output dependencies among affect dimensions, and (ii) a multi-layer hybrid framework composed of a temporal regression layer for predicting affect 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. We demonstrate the effectiveness and the robustness of the proposed frameworks by subject-independent experimental validation(s) performed on a naturalistic data set of facial expressions.*

## 1. Introduction

Traditionally, research in the field of automatic affect recognition has focused on recognizing discrete, basic emotional states (e.g. happiness, sadness) from posed data acquired in laboratory settings [3]. 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 affect where an emotional state is characterized in terms of a number of latent dimensions [13]. 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 affect prediction and recognition has increasingly attracted the interest of the affective computing researchers in recent years.

Some representative works include that of [16] quantizing the V-A dimensions into 4 or 7 levels and using Conditional Random Fields for classification from audio cues, [8] discriminating emotions into more coarse classes (such as positive vs. negative) by combining audio-visual cues via Coupled Hidden Markov Models (CHMMs) and likelihood space fusion, [17] utilizing a dynamic Bayesian network combined with Long-Short Term Memory Neural Nets (LSTM-NNs) for (quantized) quadrant prediction.

Despite such interest and progress in the field, how to design emotion-specific prediction and classification frameworks that can handle multimodal (and spontaneous) data has not yet been investigated. Kim and Andre have recently proposed a novel scheme of emotion-specific multilevel dichotomous classification (EMDC) using the property of the dichotomous categorization in the 2D emotion model (valence and arousal). They exploit the fact that arousal classification yields a higher correct classification ratio than valence classification (or direct multiclass classification) [4]. They apply this scheme on classification of four emotion classes (positive/high arousal, negative/high arousal, negative/low arousal and positive/low arousal) from physiological signals (recorded in the context of listening to music). The issue of how to create such frameworks for dimensional and continuous prediction of emotions, taking into account other modalities (e.g., vision and audio) and various aspects of emotion representation (quantized vs. continuous), remains open. This paper aims to make a contribution in this direction. Motivated by relevant psychological findings, it focuses on the design of prediction/classification

frameworks suitable for handling the compound nature of representation in the (continuous) dimensional affect space. Findings in the fields of emotion cognition and psychology suggest that the V-A dimensions are inter-correlated [1, 5, 6, 11]. Therefore, we propose two frameworks that enable the learning of such correlations and generate more substantiated and robust affect prediction/classification:

**Output-associative Relevance Vector Machine Regression (OA-RVM).** A sparse and probabilistic regression framework that extends the traditional RVM regression by being able to learn temporal output correlations (via output-association).

**Multi-layer hybrid classification (MF-Hybrid).** A framework that is composed of three distinct layers: (i) A regression layer generating the continuous A-V prediction (by using Long-Short Term Memory Neural Nets), (ii) a graphical model layer trained on the predicted affect dimensions to capture the correlations between the continuous affect descriptions (by proposing and using Auto-Regressive Coupled HMM), and (iii) a final discriminative classification and fusion layer (using Support Vector Machine) for incorporating informative statistics extracted from both the regression and the graphical model layer.

To date, no such work has been attempted for dimensional affect prediction/classification. We investigate the feasibility and the usefulness of the proposed OA-RVM and MF-Hybrid frameworks on the highly challenging problems of dimensional prediction and classification of emotions from naturalistic facial expressions. We demonstrate with experimental evaluations the robustness and the effectiveness of the proposed frameworks.

## 2. Data set

We used the Sensitive Artificial Listener (SAL) Database [2] 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 (2 female and 2 male) have been continuously annotated by 3-4 coders along the V-A dimensional (affect) space. Representative frames together with their facial point trackings are shown in Fig. 1.

Based on the annotations provided, we used a set of automatic segmentation and ground truth generation algorithms [9] to obtain segments of positive/negative emotional displays. 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 non-positive to positive and back to non-positive).

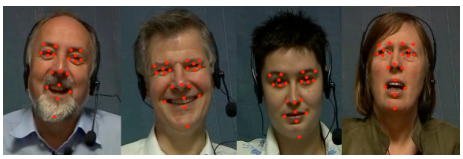

Figure 1. Examples of SAL data along with the tracked 20 points for affect prediction from facial expressions.

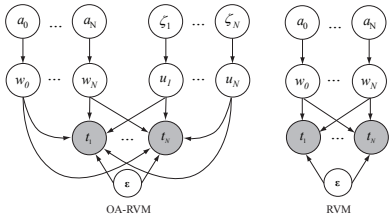


Figure 2. Graphical model comparison of OA-RVM and RVM (shaded nodes are observed variables).

## 3. Feature Extraction

For tracking the facial feature movements displayed during the naturalistic interactions, we use the tracking scheme introduced in [12]. We track the corners of the eyebrows (4 points), the eyes (8 points), nose (3 points), mouth (4 points) and chin (1 point). For each video episode containing $n$ frames, the tracker results in a feature set with dimensions $n * 20 * 2$. Fig. 1 shows example frames from the data set employed, together with the tracking of the facial feature points.

## 4. The Output Associative RVM Regression Framework

In this section, we briefly describe the two generic methods used, namely, Relevance Vector Machine (RVM) and Support Vector Machines (SVM) for Regression (i.e. SVR), and subsequently introduce the design of the OA-RVM framework for dimensional affect prediction.

### 4.1. RVM and SVM Revisited

We assume a (multidimensional) regression problem with $N$ training examples, $(\mathbf{x_i}, \mathbf{t_i})$. In the Bayesian framework applied in RVM, our goal is to learn the functional:

$$t_i = \mathbf{w}^T \phi(\mathbf{x_i}) + \epsilon_i \quad (1)$$

where the $\epsilon_i$ are assumed to be independent Gaussian samples with zero mean and $\sigma^2$ variance, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. $\phi$ is a typically non-linear projection of the input features, $\mathbf{x_i}$. The method infers the set of weights $\mathbf{w}$ along with the noise estimation, given the training data. SVR, on the other hand, employs Lagrangian optimization to determine the weights $\mathbf{w}$, with no explicit noise modeling. Structural risk minimization is applied to minimize overfitting.

### 4.2. OA-RVM

In this section we describe the proposed OA-RVM framework. Firstly, to obtain the output associative functional, we increment Eq. 1 as follows:

$$t_i = \mathbf{w}^T \phi_w(\mathbf{x_i}) + \mathbf{u}^T \phi_u(\mathbf{y_i^v}) + \epsilon_i \quad (2)$$

Where each $\mathbf{y_i^v}$ is a vector of multi-dimensional outputs over a temporal window of $[i-v, i+v]$. [1] The $\mathbf{y_i^v}$ features are called the *output features*, while $\mathbf{x}$ are called the *input features*, henceforth. Note that the output features can be estimated by predicting the multi-dimensional ground truth using any (noisy and imperfect) prediction scheme. The goal now becomes learning not only the set of weights ($\mathbf{w}$) for the input features, but also the set of weights ($\mathbf{u}$) for the output features along with the noise estimate, ($\epsilon_i$).

**The Framework.** We now specify the Bayesian framework which describes our model. Firstly, we consider $\mathbf{\Phi_w}$ ($N \times M_u$) to be the basis matrix attained by applying a selected kernel to the input features $\mathbf{x}$, and $\mathbf{\Phi_u}$ ($N \times M_w$) respectively for the output features, $\mathbf{y^v}$ (the columns, $M_u$ and $M_w$, refer to the complete set of basis vectors of dimensionality $N$). Then, by extending Eq. 2 we obtain:

$$\mathbf{t} = \Phi_w \mathbf{w} + \Phi_u \mathbf{u} + \boldsymbol{\epsilon} = \Phi_{wu} \mathbf{w_u} + \boldsymbol{\epsilon} \quad (3)$$

where $\mathbf{\Phi_{wu}} = [\mathbf{\Phi_w} | \mathbf{\Phi_u}]$ is an $N \times (M_u + M_w)$ matrix and $\mathbf{w_u} = [\mathbf{w_1} \dots \mathbf{w_{M_w}} | \mathbf{u_1} \dots \mathbf{u_{M_u}}]^T$ is the concatenated vector of weights. Thus, the complete data set likelihood is formulated as:

$$P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \sigma^2) = \prod_{i=1}^{N} N(\mathbf{w_u}^T [\phi_w(\mathbf{x_i}) | \phi_u(\mathbf{y_i^v})], \sigma^2)$$

Following the Bayesian approach of RVM [14], we need to set the hyperpriors on our weights. Each set of weights ($\mathbf{w}, \mathbf{u}$) is assigned a Gaussian zero-mean prior to express preference over smaller weights, thus infer smoother, less complex functions and induce sparsity:

$$P(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{M_u} \mathcal{N}(0, \alpha_i^{-1}), P(\mathbf{u}|\boldsymbol{\zeta}) = \prod_{i=1}^{M_w} \mathcal{N}(0, \zeta_i^{-1}) \quad (4)$$

We have now introduced two vectors of hyperparameters, $\boldsymbol{\alpha}$ (as originally used in RVM) and $\boldsymbol{\zeta}$ (for our output features), each controlling the distribution of each of the weights.

**Inference.** The goal of the inference procedure is to infer the unknown parameters of our problem given the training data. The posterior is decomposed as:

$$P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2 | \mathbf{t}) = \frac{P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2)}{p(\mathbf{t})} \quad (5)$$

[1] For frame based online application, we can limit the context to past input only, i.e. $[i-v, i]$. Futhermore, the output window regards *only* the output dimensions since we study the effect of output-covariances.

Ideally, given a new test data $x_*$, we would like to predict target $t_*$ by estimating $p(t_*|\mathbf{t})$:

$$\int P(\mathbf{t_*}|\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2 | \mathbf{t}) d\mathbf{w} d\mathbf{u} d\boldsymbol{\alpha} d\boldsymbol{\zeta} d\sigma^2 \quad (6)$$

Unfortunately, a direct estimation is intractable, thus an approximation is employed. Similarly to the original RVM formulation [14], we decompose the posterior as follows:

$$P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2 | \mathbf{t}) = P(\mathbf{w}, \mathbf{u} | \mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2 | \mathbf{t}) \quad (7)$$

Using the Bayes theorem we obtain:

$$P(\mathbf{w}, \mathbf{u} | \mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) = \frac{P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \sigma^2) P(\mathbf{w}, \mathbf{u}|\boldsymbol{\alpha}, \boldsymbol{\zeta})}{P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2)} \quad (8)$$

This calculation is tractable, since all components are Gaussian distributions and it is well known that products and divisions of Gaussian distributions result also in Gaussian distributions. We will firstly examine the joint probability. By assuming independence, we obtain $P(\mathbf{w}, \mathbf{u}|\boldsymbol{\alpha}, \boldsymbol{\zeta})$, a zero-mean Gaussian with a covariance matrix $\mathbf{A_Z} = \text{diag}(\alpha_1 \dots \alpha_{M_w}, \zeta_1 \dots \zeta_{M_u})$.

$$P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) = \int P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \sigma^2) P(\mathbf{w}, \mathbf{u}|\boldsymbol{\alpha}, \boldsymbol{\zeta}) d\mathbf{w} d\mathbf{u} \quad (9)$$

is a convolution of Gaussians and after replacing with the defined variables $\mathbf{w_u}$, $\mathbf{A_z}$ and $\mathbf{\Phi_{wu}}$, it is shown [14] to be a zero-mean Gaussian distribution with covariance matrix $\sigma^2 \mathbf{I} + \Phi_{wu} \mathbf{A_Z^{-1}} \Phi_{wu}^T$. Finally, Eq. 8 is considered to be a Gaussian distribution with a mean $\boldsymbol{\mu} = \sigma^2 \Sigma \mathbf{\Phi}_{wu}^T \mathbf{t}$ and a covariance matrix $\boldsymbol{\Sigma} = (\mathbf{A_Z} + \sigma^2 \mathbf{\Phi}_{wu}^T \mathbf{\Phi}_{wu})^{-1}$. Returning to the second component $P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2 | \mathbf{t})$ of the posterior in Eq. 7, by following the Bayes rule, we find it to be proportional to:

$$P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2 | \mathbf{t}) \propto P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) P(\boldsymbol{\alpha}) P(\boldsymbol{\zeta}) P(\sigma^2) \quad (10)$$

By assuming uniform uninformative hyperpriors [14], we need to maximize $P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2)$ with respect to the hyperparameters. Again, we have a convolution of Gaussians (Eq. 9) which in turn generates another zero mean Gaussian with covariance matrix $\sigma^2 \mathbf{I} + \mathbf{\Phi}_{wu} \mathbf{K}^{-1} \mathbf{\Phi}_{wu}^T$. The maximization of this probability can be performed by expectation maximization as described in [14] or the faster marginal maximization algorithm proposed in [15]. The most probable values ($MP$) are selected by the chosen optimization procedure (e.g., [14, 15]). We adopt an approximation of $P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2 | \mathbf{t})$ in Eq. 7 by replacing it with a delta function at its mode.

**Prediction.** Given a new (multi-dimensional) input data $\mathbf{x}_*, \mathbf{y}_*^\mathbf{v}$, we want to calculate $t_*$ given the training data. By considering $\boldsymbol{\alpha_z} = [a_1 \dots a_{M_w}, \zeta_1 \dots \zeta_{M_u}]$ and using Eq. 6 and Eq. 8 we obtain:

$$P(t_*|\mathbf{t}, \boldsymbol{\alpha}_{\mathbf{z}MP}, \sigma^2_{MP}) =$$

$$\int P(t_*|\mathbf{w_u}, \sigma^2_{MP}) P(\mathbf{w_u}|\mathbf{t}, \boldsymbol{\alpha}_{\mathbf{zMP}}, \sigma^2_{MP}) d\mathbf{w_u} \quad (11)$$

Again, this is a convolution of Gaussians and it can be shown that

$$P(t_*|\mathbf{t}, \boldsymbol{\alpha}_{\mathbf{z}MP}, \sigma^2_{MP}) \sim N(t_*|\sigma^2_*) \quad (12)$$

where

$$t_* = \boldsymbol{\mu}^T_{wu}[\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}^{\mathbf{v}}_*)] \quad (13)$$

$$\sigma^2_* = \sigma^2_{MP} + [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}^{\mathbf{v}}_*)]^T \boldsymbol{\Sigma} [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}^{\mathbf{v}}_*)] \quad (14)$$

with variance $\sigma^2_*$ (which relates to the confidence in our prediction). The parameter vector $\boldsymbol{\mu}_{wu}$ contains the weights for the input and output relevance vectors, i.e. $\boldsymbol{\mu}_{wu} = [\boldsymbol{\mu}_w|\boldsymbol{\mu}_u]$. The basis matrix for a new set of test points should now contain both the distances from the new test input features $\mathbf{x}_*$ to all the input feature relevance vectors, as well as the test output feature $\mathbf{y}^{\mathbf{v}}_*$ distances to the output feature relevance vectors. The graphical models of both OA-RVM and RVM are illustrated in Fig. 2. For further details on the framework, as well as more extensive experiments and results, the readers are referred to [10].

## 5. The Multi-layer Hybrid Framework

The MF-Hybrid is designed as a three-layer hybrid framework for dimensional affect classification. MF-Hybrid incorporates a prediction layer for each emotion dimension, a graphical model layer for capturing inter-dimensional emotion correlations, and a final (discriminative) classification and fusion 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.

### 5.1. Long-Short Term Memory Neural Nets

LSTM Neural Nets (LSTM-NN) 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.

### 5.2. Auto-Regressive Coupled HMM

By merging two common HMM variants, namely the Coupled HMM and the Auto-Regressive HMM, we propose the design of 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 the typical (two) hidden nodes as well as two non-hidden nodes modeling each observation stream. The state of each hidden node at time $t$ depends on the hidden states of both hidden nodes at $t-1$. Auto-regression in HMMs (ACHMM) relaxes the assumption that the observation depends only on the current state.

It models the distribution of an observation at time $t$ conditioned on the current hidden state as well as the previous observation. An ACHMM with two streams of observations is illustrated in Fig. 3(a) (see the HMM nodes).

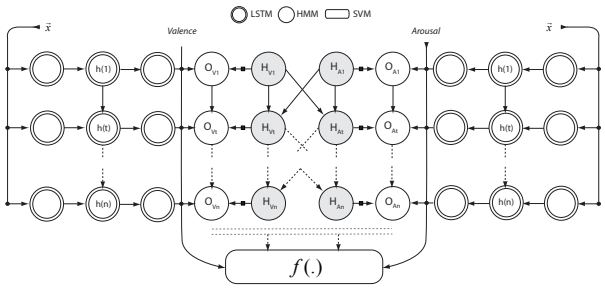


Figure 3. Illustration of the proposed (3-layer) MF-Hybrid. The graphical model employed is an ACHMM, where shaded nodes represent hidden nodes and filled squares represent a mixture of Gaussians.

### 5.3. MF-Hybrid

The multi-layer hybrid framework we propose for dimensional affect classification is illustrated in Fig. 3. The framework has three distinct layers: (1) a regression layer (LSTM-NNs) which generates the continuous prediction for each affect dimension, (2) a graphical model layer (ACHMM) that models the inter-dimensional structure, and (3) finally, a (discriminative) classification and fusion 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 affect dimension (valence/arousal). The graphical model layer on the other hand, can capture subtle, emerging inter-dimensional patterns which seamlessly contribute to the dimensional affect classification of a given sequence.

Let us consider the **LSTM-ACHMM$_{\mathbf{ML}}$** model. Firstly, two separate LSTM-NNs are trained as regressors, one for each affect dimension. Let $\mathcal{D}$={Valence,Arousal} be the set

of affect 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 affect dimension $d \in \mathcal{D}$ depends directly on the previous observation for this dimension $O_{dt-1}$, and the state of the hidden node $H_{dt}$. Moreover, the hidden state $H_{dt}$ depends on the previous hidden states of both dimensions ($H_{At}$ and $H_{Vt}$) due to the coupled nature of the model. Classification is obtained using the Maximum Likelihood (ML) principle (classifying the entire episode based on the model that produces the maximum likelihood).

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* via 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**$_{\text{SVM}}$. 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. $\mathbf{\bar{MPE}_d(s)}$ (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 time frame 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 affect dimensions are fed into the classifier, thus enabling more robust learning of inter-dimensional patterns and structure.

# 6. Experiments and Results

We conducted a set of experiments in order to validate the proposed OA-RVM regression and the MF-Hybrid classification frameworks, separately. The following sections provide details on the experimental setups adopted, experiments conducted and results obtained.

## 6.1. Evaluation of OA-RVM

**Experimental setup.** We use the traditional RVM as the baseline for our comparisons with OA-RVM. We also use SVR as it is one of the most widely adopted regression techniques in the field. The kernel used for the construction of the basis matrices is a Gaussian, $K(x, x_i) = exp\left\{(-(x - x_i)^2)/r^2\right\}$ where $r$ stands for the width of the function. The window parameter $v$ in the output-associative functional we employ is generally varied in the range of $[0, 18]$ and can be determined by cross-validation. It should be noted that for the probabilistic regression methods (RVM, OA-RVM), the hyperparameters are determined by optimizing the likelihood function (by using the fast marginal likelihood maximization algorithm proposed in [15]). We use RVM to obtain the initial output estimation (i.e., the output features) for OA-RVM. For SVR we apply cross-validation employing an $\epsilon$-insensitive loss function. In our current setting, we assume that the episodes contained in our data set have been coarsely classified into either positive or negative, prior to the prediction (regression) procedure. This assumption is motivated by the fact that we would like to focus on the prediction results in more detail, and study them in isolation for each class (e.g., which dimension is easier to predict for which class). Based on the aforementioned assumptions, we conduct subject-independent experiments by using data from one subject *only* for training, and subsequently using the data from the remaining three subjects for testing. We evaluate the proposed model in terms prediction accuracy using the root mean squared error (RMSE) that incorporates the bias and variance of the prediction.

**Results.** Table 1 presents the subject-independent prediction results in terms of RMSE and window size (v) employed. Each row on the table presents the results obtained by training the model using data from one subject (indicated in the first column) and testing data from the rest of the subjects. OA-RVM provides better prediction results than RVM and SVR, for each and every tested case. Overall, valence appears to be easier to predict than arousal for the negatively valenced emotions, while arousal appears to be easier to predict for the positively valenced emotions. The maximum output-associative window size of $v = 18$ appears to provide the best prediction results in many cases, while on average, a window of size $v > 9$ appears to be optimal. Overall, naturalistic emotional expressions are highly subject-dependent [3]. Yet, our experimental results

indicate that automatic, subject-independent, dimensional and continuous prediction of emotions becomes feasible by utilizing input and output associations as well as temporal context.

Table 1. Subject-independent prediction results (RMSE) for SVR, RVM and OA-RVM .

| | Valence | | | | Arousal | | | |
|---|---|---|---|---|---|---|---|---|
| **POS** | **SVR** | **RVM** | **OA-RVM** | **v** | **SVR** | **RVM** | **OA-RVM** | **v** |
| subj1 | 0.21 | 0.16 | 0.15 | 18 | 0.16 | 0.16 | 0.15 | 18 |
| subj2 | 0.22 | 0.26 | 0.17 | 18 | 0.18 | 0.18 | 0.14 | 9 |
| subj3 | 0.22 | 0.22 | 0.22 | 12 | 0.17 | 0.17 | 0.16 | 12 |
| subj4 | 0.19 | 0.16 | 0.15 | 6 | 0.19 | 0.14 | 0.13 | 18 |
| **NEG** | **SVR** | **RVM** | **OA-RVM** | **v** | **SVR** | **RVM** | **OA-RVM** | **v** |
| subj1 | 0.11 | 0.10 | 0.09 | 12 | 0.36 | 0.39 | 0.35 | 18 |
| subj2 | 0.14 | 0.11 | 0.09 | 14 | 0.37 | 0.33 | 0.32 | 10 |
| subj3 | 0.10 | 0.10 | 0.10 | 5 | 0.37 | 0.40 | 0.37 | 18 |
| subj4 | 0.13 | 0.11 | 0.09 | 18 | 0.14 | 0.13 | 0.13 | 2 |

### 6.2. Evaluation of MF-Hybrid

**Experimental setup.** For our experiments, we use the bidirectional LSTM-NNs with one hidden layer. The ACHMMs have 3 hidden states for each hidden node. 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*, using data from three subjects for training and using the data from the remaining subject *only* for testing.

**Results.** The results obtained are shown in Table 2. The LSTM-NN 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 counterpart 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%). In summary, the proposed framework provides an increase in accuracy in all classification tasks. Specifically, our results indicate that by modeling inter-dimensional covariances we can learn complex and subtle class variances more robustly.

## 7. Conclusions

Findings in the fields of emotion cognition and psychology suggest that the V-A dimensions are inter-correlated [1, 5, 6, 11]. Motivated by such findings, this paper focused on designing prediction/classification frameworks suitable for handling the compound nature of affect representation in (continuous) dimensional space. More specifically, it introduced two affect prediction/classification frameworks: (i) an output-associative Relevance Vector Machine Regression framework (OA-RVM) for continuous affect prediction, and (ii) a multi-layer hybrid classification framework (MF-Hybrid) for V-A hemispheric and quadrant classification. The Output-Associative Relevance Vector Machine (OA-RVM) regression framework augments the traditional RVM by being able to learn *non-linear input-output dependencies*. Instead of depending solely on input patterns, OA-RVM models output structure and covariances within a predefined temporal window, thus capturing past and future context. The Multi-layer Hybrid Framework is composed of a regression layer (using LSTM-NN) which generates the continuous prediction for each affect dimension, a graphical model layer (introducing and using ACHMM) that models the inter-dimensional structure, and a discriminative classification and fusion layer (using SVM) incorporating statistics from the lower layers.

We successfully applied the proposed frameworks for subject-independent dimensional affect prediction/classification from facial expressions, and demonstrated their respective advantages and efficiencies over a set of experiments. Our results show that OA-RVM outperforms both RVM and SVR in terms of prediction accuracy. Employing a temporal (output) window, which induces the learning of past and future context, contributes significantly to the prediction accuracy. Our results also show that designing a multi-layer hybrid framework (e.g., by combining LSTM-NN, Auto-Regressive CHMM, and SVM) combines the advantages of various predictors and classifiers, and provides an increase in accuracy and robustness for the valence/arousal and quadrant classification tasks. The prediction and classification frameworks introduced in this paper have been treated separately, without attempting to use one framework as part of the other one. As future work, the proposed frameworks remain to be linked, and evaluated on databases with a larger number of subjects (e.g., SEMAINE [7]) in order to obtain deeper insights into the accuracy improvement provided by these models.

Table 2. Experimental results (ACC: accuracy, PREC: precision, REC: recall) for valence (positive/negative), arousal (active/inactive) and quadrant classification.

| Layer | 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$_{ML}$** | 72 | 76 | 70 | 69 | 75 | 67 | 60 | 58 | 67 | 66 | 66 | 66 |
| **layers 1, 2, 3** | **LSTM-ACHMM$_{SVM}$** | 86 | 87 | 86 | **86** | 84 | 83 | 79 | **79** | 84 | 90 | 75 | **77** |

Overall, although the proposed frameworks have been applied to emotion-specific prediction and recognition problems, due to their highly flexible nature they can easily be extended and applied to other multi-dimensional (or multi-modal) prediction and classification problems.

## 8. Acknowledgments

This work has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB) and the European Communitys 7th Framework Programme [FP7/2007-2013] under the grant agreement no 231287 (SSPNet).

## References

[1] N. Alvarado. Arousal and valence in the direct scaling of emotional response to film clips. *Motivation & Emotion*, 21:323–348, 1997. 21, 25

[2] E. Douglas-Cowie and et al. The humaine database: addressing the needs of the affective computing community. In *Proc. of ACII*, pages 488–500, 2007. 21

[3] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *Int. Journal of Synthetic Emotions*, 1(1):68–99, 2010. 20, 24

[4] J. Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *IEEE Tran. on PAMI*, 30(12):2067–2083, 2008. 20

[5] R. Lane and L. Nadel. *Cognitive Neuroscience of Emotion*. Oxford Univ. Press, 2000. 21, 25

[6] P. A. Lewis and et al. Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, 17(3):742–748, Mar 2007. 21, 25

[7] G. McKeown, M. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Proc. of IEEE ICME*, pages 1079–1084, July 2010. 25

[8] M. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *Proc. of ICPR*, pages 3695–3699, 2010. 20

[9] M. Nicolaou, H. Gunes, and M. Pantic. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In *Proc. of LREC Int. Workshop on Multimodal Corpora*, pages 43–48, 2010. 21

[10] M. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2011. 23

[11] A. M. Oliveira and et al. Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity. In *Proc. of Annual Meeting of Int. Society for Psychophysics*, pages 245–250, 2006. 21, 25

[12] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Proc. of IEEE AFGR*, pages 97–102, 2004. 21

[13] J. A. Russell. A circumplex model of affect. *Journal of Personality & Social Psychology*, 39:1161–1178, 1980. 20

[14] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, 2001. 22

[15] M. E. Tipping and A. Faul. Fast marginal likelihood maximisation for sparse bayesian models. In *Proc. of Int. Workshop on AI and Statistics*, pages 3–6, 2003. 22, 24

[16] M. Wollmer et al. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. of Interspeech*, pages 597–600, 2008. 20

[17] 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. 20