Output-Associative RVM Regression for Dimensional and Continuous Emotion Prediction

Mihalis A. Nicolaou, Hatice Gunes and Maja Pantic

***Abstract*— Many problems in machine learning and computer vision consist of predicting multi-dimensional output vectors given a specific set of input features. In many of these problems, there exist inherent** *temporal and spacial dependencies* **between the output vectors, as well as *repeating output patterns* and *input-output associations*, that can provide more robust and accurate predictors when modelled properly. With this intrinsic motivation, we propose a novel 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*. Instead of depending solely on the input patterns, OA-RVM models output structure and covariances within a predefined temporal window, thus capturing past, current and future context. As a result, output patterns manifested in the training data are captured within a formal probabilistic framework, and subsequently used during inference. As a proof of concept, we target the highly challenging problem of *dimensional and continuous prediction* of emotions from naturalistic facial expressions. We demonstrate the advantages of the proposed OA-RVM regression by performing both subject-dependent and subject-independent experiments using the SAL database. The experimental results show that OA-RVM regression outperforms the traditional RVM and SVM regression approaches in prediction accuracy, generating more robust and accurate models.**

I. INTRODUCTION

Kernel methods such as Support Vector Machines (SVM), Relevance Vector Machines (RVM) and Gaussian Processes (GP) are amongst the most dominant techniques used in machine learning and computer vision. Many problems in these fields are inherently related to the prediction of multi-dimensional, inter-correlated structured outputs (e.g. pose normalisation, pose estimation). While most machine learning techniques aim at capturing input relationships and patterns (e.g. extracted features), many problems expose an inherent dependency amongst the output dimensions (e.g. emotion dimensions). Not being able to learn such co-occurrences can result in less robust and less accurate predictors, that will not be able to exploit specific output configurations manifested in the training data.

With these intrinsic motivations, we introduce the output-associative RVM (OA-RVM) regression, a framework that extends the traditional RVM regression by being able to learn temporal output correlations. As we show by means of various experiments, OA-RVM appears to be advantageous against traditional RVM not only in terms of prediction accuracy but also in terms of sparsity of the final model (i.e., dependence on a small number of basis vectors), thus resulting in a simpler and more robust model. To evaluate the proposed technique, we apply it to a highly challenging and suitable problem: dimensional and continuous emotion prediction.

Most research in automatic emotion recognition and prediction has focused on examining posed data acquired in laboratory settings [1], [2] in terms of basic emotional states (e.g., happiness, sadness, surprise). However, many studies show that in everyday life interactions, humans exhibit subtle affective states that do not fall under the basic emotional states (e.g. bored or interested). In order to represent and model such states, a dimensional and continuous description of human affect is employed, where an affective state can be described by a number of latent dimensions [3]. We focus on the two dimensions which are considered to cover most of the affect variability [4]: The valence dimension (V) which describes how positive or negative an emotional state is, and the arousal dimension (A) which relates to how excited or apathetic an emotional state is [5].

Our motivation for the work presented in this paper is three-fold. Firstly, dimensional and continuous affect prediction (as opposed to discrete and quantised recognition) and output-associative structured prediction are two highly inter-related problems. Psychological evidence has shown that the V-A dimensions are inter-correlated [4], [6]–[8]. Therefore, the proposed scheme aims to enable the learning of such correlations and generate more substantiated predictions by embedding in the model an initial output estimation (using RVM) together with the original input features. Secondly, temporal dynamics play a significant role in emotion recognition [1], [2]. The proposed OA-RVM regression aims to capture the temporal dynamics by employing a temporal window (covering a set of past and future outputs) in order to accommodate temporal (output) patterns both in past and future context. Thirdly, dimensional and continuous prediction of emotions is a relatively unexplored area in the field of affective computing, and which prediction method is best suited to the task is still unknown. Therefore, as well as validating the proposed OA-RVM model with comprehensive experiments, we also compare it to traditional regression techniques such as RVM and Support Vector Regression (SVR). In the following, we briefly review related work on output-associative structured regression and dimensional and continuous emotion prediction, and subsequently list the contributions of our work.

Output-Associative Structured Regression: Output-associative structured regression has gained much popularity

The authors are with Department of Computing, Imperial College London, U.K. { mihalis, h.gunes, m.pantic }@imperial.ac.uk. M. Pantic is also with the Faculty of EEMCS, University of Twente, The Netherlands

over the last years within the pattern recognition community. Kernel Dependency Estimation (KDE) was proposed in 2002 by Weston [9], with a goal of learning output dependencies using Kernel Principle Component Analysis (KPCA) and ridge regression. KDE was reformulated in 2005 by Cortes et al. [10] discarding the need for KPCA and adopting the optimisation of a cost function. KDE has been applied to problems such as string matching and image reconstruction. Previous efforts on modeling input and output covariances have motivated the extension of models such as Kernel Ridge Regression (KRR), SVM for regression [11] and GP [12]. [11] optimises an output-associative functional which incorporates outputs and inputs using primal/dual formulations and adapts the model to KRR and SVR. [12] develops the Twin GP model, which employes GP priors to model input and output relations. The Kullback−Leibler divergence is applied on the input and output distributions. Subsequently, the output targets are estimated by the minimisation of the KL divergence. Both works have been applied to modeling human pose estimation.

We choose to extend RVM as it is considered more efficient than GP [12]. Compared to the models presented in [11], [12] we offer a specific output temporal window parameter for fine-tuning our model. Furthermore, compared to [11], our OA-RVM regression framework offers a probabilistic formulation of the output-associative function by following the original RVM framework and providing explicit noise modelling.

Dimensional and Continuous Emotion Prediction: Past work on dimensional affect recognition was based on classifying emotional states by quantising the real values, into coarse binary categories of positive vs. negative [13], into quadrants of the V-A space [14] or into dense quantised levels (e.g. 7 levels [15]). [16] fuses facial expression and audio cues exploiting SVM for regression (SVR) and late fusion, using weighted linear combinations, and uses discretised annotations (on a 5-point scale, for each dimension). The works that focused on predicting continuous and real values are few. Using speech features, [15] employs recurrent neural networks (Long Short-Term Memory) and SVR, while [16] uses SVR, k-NN and a fuzzy logic estimator. None of these works have explored input-output associations and spatio-temporal dependencies between the output vectors for dimensional and continuous emotion prediction.

Contributions: Based on the aforementioned literature review, and to the best of our knowledge, this paper presents the first approach in the affective computing field that utilises input-output associations for dimensional and continuous prediction of emotions. More specifically, our work (i) proposes a novel, sparse and probabilistic regression model with output-association (OA-RVM, henceforth), taking advantage of the traditional RVM framework, and (ii) investigates the feasibility and the usefulness of the proposed OA-RVM framework on the highly challenging problem of dimensional and continuous prediction of emotions from naturalistic facial expressions.

The rest of the paper is organised as follows. In Section II, we briefly revisit the RVM and SVM models in order to provide a basis for OA-RVM, introduced and explained in Section III. Section IV describes the data set employed in our experiments, as well as the feature extraction and tracking process. Section V explains the experimental settings employed. Section VI provides a demonstration of the behaviour of the model on learning dimensional emotion annotations, while Section VII presents the experiments and and discusses the results. Finally, Section VIII concludes the paper.

II. RVM AND SVM REVISITED

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).

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:

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

where the ϵ_i are assumed to be independent Gaussian samples with zero mean and σ^2 variance, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. ϕ 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.

In the SVR, the functional $t_i = \mathbf{w}^T \phi(\mathbf{x}_i) + b$ is learnt, where ϕ is an implicit mapping to a kernel space, $\mathbf{w}$ represents the set of weights and b the bias. Lagrangian optimisation is employed to determine the optimal parameters provide the final model. In contrast to Bayesian regression methods, there is no explicit noise modelling in SVR while the structural risk minimisation principle is applied to minimize the risk of overfitting.

III. OUTPUT-ASSOCIATIVE RVM REGRESSION

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

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

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, (ϵ_i)[2].

[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.

[2]Note that in the output-associative formulation, the noise component can now be considered as the sum of the noise generated by the input features $\sigma_{\mathbf{x}}$ and the output features $\sigma_{\mathbf{y^v}}$, i.e. $\epsilon_i \sim N(0, \sigma_y^2 + \sigma_x^2) = N(0, \sigma^2)$.

A. The Framework

In this section we specify the Bayesian framework which describes our model. Firstly, we consider $\mathbf{\Phi_w}$ (N×M_u) to be the basis matrix attained by applying a selected kernel to the input features $\mathbf{x}$, and $\mathbf{\Phi_u}$ (N×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 though usually both are of dimensionality N). Then, by extending Eq. 2 we obtain:

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

where $\mathbf{\Phi_{wu}} = [\mathbf{\Phi_w}|\mathbf{\Phi_u}]$ is an N×$(M_u + M_w)$ 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:

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

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

Following the Bayesian approach of RVM [17], 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:

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

$$P(\mathbf{u}|\boldsymbol{\zeta}) = \prod_{i=1}^{M_w} \mathcal{N}(0, \zeta_i^{-1}) \tag{5}$$

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.

B. Inference

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

$$P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2|\mathbf{t}) = \frac{P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2)}{p(\mathbf{t})} \tag{6}$$

Ideally, given a new test data x_*, we would like to predict target t_*:

$$p(t_*|\mathbf{t}) =$$

$$\int P(\mathbf{t_*}|\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2|\mathbf{t}) d\mathbf{w} d\mathbf{u} d\boldsymbol{\alpha} d\boldsymbol{\zeta} d\sigma^2 \tag{7}$$

Unfortunately, the above equation is intractable, thus an approximation is needed. Therefore, similarly to the original RVM formulation [17], we decompose the posterior as follows:

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

Using the Bayes theorem we obtain:

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

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 distribution with a covariance matrix $\mathbf{A_Z}$ = diag$(\alpha_1 \dots \alpha_{M_w}, \zeta_1 \dots \zeta_{M_u})$.

$$P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) = \int P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \sigma^2) P(\mathbf{w}, \mathbf{u}|\boldsymbol{\alpha}, \boldsymbol{\zeta}) d\mathbf{w} d\mathbf{u} \tag{10}$$

is a convolution of Gaussian and after replacing with the defined variables $\mathbf{w_u}$, $\mathbf{A_z}$ and $\mathbf{\Phi_{wu}}$, it is shown [17] 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. 9 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 $\mathbf{\Sigma} = (\mathbf{A_Z} + \sigma^2 \mathbf{\Phi}_{wu}^T \mathbf{\Phi}_{wu})^{-1}$.

Returning to the second component $P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2|\mathbf{t})$ of the posterior in Eq. 8, by following the Bayes rule, we find it to be proportional to:

$$P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2|\mathbf{t}) \propto P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2) P(\boldsymbol{\alpha}) P(\boldsymbol{\zeta}) P(\sigma^2) \tag{11}$$

By assuming uniform uninformative hyperpriors [17], we need to maximise $P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2)$ with respect to the hyperparameters. Again, we have a convolution of Gaussians (Eq. 10) which in turn generates another zero mean Gaussian distribution with covariance matrix $\sigma^2 \mathbf{I} + \mathbf{\Phi}_{wu} \mathbf{K}^{-1} \mathbf{\Phi}_{wu}^T$. The maximisation of this probability can be performed by expectation maximisation as described in [17] or the faster marginal maximisation algorithm proposed in [18]. The most probable values (MP) are selected by the chosen optimisation procedure ([17], [18]), while we adopt an approximation of $P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2|\mathbf{t})$ in Eq. 8 by replacing it with a delta function at its mode.

C. 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. 7 and Eq. 9 we obtain:

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

$$\int P(t_*|\mathbf{w_u}, \sigma_{MP}^2) P(\mathbf{w_u}|\mathbf{t}, \boldsymbol{\alpha_{zMP}}, \sigma_{MP}^2) d\mathbf{w_u} \tag{12}$$

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

$$P(t_*|\mathbf{t}, \boldsymbol{\alpha}_{\mathbf{z}MP}, \sigma_{MP}^2) \sim N(t_*|\sigma_*^2) \tag{13}$$

where

$$t_* = \boldsymbol{\mu}_{wu}^T [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^\mathbf{v})] \tag{14}$$

$$\sigma_*^2 = \sigma_{MP}^2 + [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^\mathbf{v})]^T \mathbf{\Sigma} [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^\mathbf{v})] \tag{15}$$

with the variance σ_*^2 (which relates to the confidence in our prediction). The parameter vector $\boldsymbol{\mu}_{wu}$ contains the weights

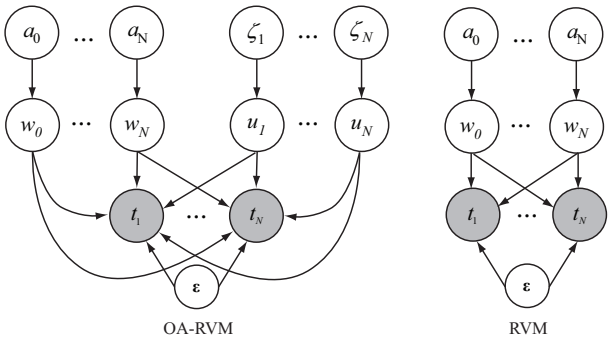

Fig. 1. Graphical model comparison of RVM and OA-RVM. Shaded nodes are observed variables.

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. 1.

D. Complexity

The parameter determination algorithm of RVM generally involves the optimisation of a non-convex function. The basis matrix for RVM is considered to be N×M, with M basis functions. An inversion of this matrix is required, which induces $O(M^3)$ computational complexity. In OA-RVM, without loss of generality, we assume that we have a N×2M basis matrix: A dimensionality of M for the input features and an additional M for the output features. Thus, the complexity is $O((2M)^3) = O(M^3)$. Furthermore, to obtain the output features for OA-RVM we apply the original RVM algorithm. If for a d-dimensional output problem, the complexity of the original RVM algorithm is $O(dC)$, then for OA-RVM the complexity would be $2O(dC)$ which is still $O(dC)$. Therefore, OA-RVM induces no further computational complexity to the RVM algorithm.

IV. DATA SET AND FEATURE EXTRACTION

As a proof of concept for this work, we use the Sensitive Artificial Listener (SAL) Database [19]. It contains audio-visual, naturalistic affective conversational data taking place between a participant and an avatar (operated by a human). Each avatar is considered to have a different personality: Poppy is happy, Obadiah is gloomy, Spike is angry and Prudence is pragmatic.

The recordings were made in a controlled laboratory setting with one camera, microphones, uniform background and constant lighting conditions. As our aim is to achieve continuous emotion prediction, we could only take advantage of the amount of data which was annotated in the *valence-arousal dimensional affect space*. This corresponds to a portion of the database that contains data from 4 subjects (subjects 1 and 2 are female, and subjects 3 and 4 are male) and their respective annotations (provided by 3-4 coders).

Fig. 2. Examples of the data at hand from the SAL database along with the extracted 20 points, used as features for the facial expression cues.

Frames from this portion of the SAL database, together with the trackings of facial points, are shown in Fig. 2. Based on the annotations provided, we used a set of automatic segmentation and ground truth generation algorithms [20] that generated segments of positive/negative emotional displays. More specifically, we generated segments capturing transitions to an emotional state and back (e.g., going from non-positive to positive and back to non-positive). Henceforth, we refer to these classes as positive for the transition to a positive emotional state, and negative for the transition to a negative emotional state. In total, we used 61 positive and 73 negative segments, and approximately 30,000 video frames.

For feature extraction, we employ the Patras - Pantic particle filtering tracking scheme [21] for tracking the facial feature movements displayed during the naturalistic interactions. 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 segment containing n frames, the tracker results in a feature set with dimensions $n * 20 * 2$. Fig. 2 shows examples from the data set employed together with the tracking of the facial feature points.

V. EXPERIMENTAL SETTING

We conducted comprehensive experiments in order to validate the proposed OA-RVM regression framework, and investigate its feasibility and usefulness for dimensional and continuous prediction of emotions.

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 (Eq. 1) is generally varied in the range [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 optimising the likelihood function (by using fast marginal likelihood maximisation algorithm proposed in [18]). 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 ϵ-insensitive loss function.

In our current setting, we assume that the segments contained in our data set (Section IV) have been coarsely classified into either positive or negative, prior to the prediction (regression) procedure. The classification stage is beyond the scope of this paper, and can be achieved by applying an accurate (coarse) classifier, e.g. [13], on top of the current scheme. 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 two types of experiments.

Subject-dependent experiments. For each subject we divide the data into equal training and testing sets. We predict the emotional dimensions for each subject separately over 2-fold cross-validation. We present the average of these results.

Subject-independent experiments. Subject-independent experiments are generally considered difficult when data from only a few subjects are available [15]. We conduct subject-independent experiments in a more challenging scenario where we use the data from one subject *only* for training, and subsequently use the data from the remaining three subjects for testing.

We evaluate our models in terms of both prediction accuracy and sparsity. For prediction accuracy, we employ the root mean squared error (RMSE) estimation that incorporates the bias and variance of the prediction. To evaluate sparsity, we refer to the number of relevance vectors (RVs) retained by the model after training (for RVM and OA-RVM). While evaluating the sparsity of OA-RVM, we consider the output features (the initial output estimation provided via RVM) as part of the initialisation, and thus evaluate the sparsity of the final model. Since these RVs correspond to basis vectors centered on a training example, we can infer which and how many training examples are considered significant and retained for the specific task at hand. A smaller set of RV implies a less complex model, with a reduced risk of overfitting.

VI. WHY OUTPUT-ASSOCIATION FOR CONTINUOUS EMOTION PREDICTION?

In this section, we would like to demonstrate how the proposed OA-RVM regression framework is efficiently applicable to the problem of automatic emotion prediction in a continuous dimensional space. We focus our analysis and discussion on Fig. 3. The figure illustrates how employing the original RVM and the proposed OA-RVM provides continuous prediction of valence and arousal dimensions for one training sequence (consisting of 315 frames) extracted as explained in Section IV.

The predictions generated by RVM are shown in Fig. 3(a,b) while the OA-RVM generated predictions with a window of $v = 0$ and $v = 4$ are shown in Fig. 3(c,d) and Fig. 3(e,f), respectively. The ground truth for both the valence and the arousal dimensions is shown in all figures as *gTruth*, for comparison. The generated predictions for valence appear on the left column of Fig. 3, while the generated predictions for arousal appear on the right. The window of $v = 0$ is meant to represent the most sparse results, while a window of $v = 4$ is deemed sufficient for a sequence of 315 frames as it embeds 9 temporal steps (frames) in terms of past (4 frames), present (current frame) and future (4 frames) context.

In this particular sequence, the subject appears to be displaying negatively valenced emotions (e.g., sadness, disappointment), with a decreasing arousal over time (towards a more passive emotional state). In the figure we observe how the RVM framework generates predictions (depicted with RVM line) by using 32 relevance vectors (RVs) for valence (Fig. 3a) and 39 RVs for arousal (Fig. 3b). Fig. 3(c,d) then illustrates how the proposed OA-RVM framework generates predictions for the sequence at hand, for valence and arousal, with a temporal window of $v = 0$. Note how OA-RVM is able to learn a *smoother* and *more accurate* model by using just 7 RVs for valence and 6 RVs for arousal, respectively.

As specified in Eq. 2, OA-RVM depends on both the input features ($\mathbf{x}$, depicted as *IF* in the figure) and the output features ($\mathbf{y^v}$, depicted as *OF* in the figure). To illustrate the behaviour of the framework, we decompose the relevance vectors (RVs) selected by OA-RVM into the RVs centred around the input features (RV-IF) and the RVs centred around the output features (RV-OF).

For the valence dimension, the 7 RVs used for the OA-RVM model can be decomposed into 4 RVs corresponding to input features (the relevant frames shown in Fig. 3c) and 3 RVs corresponding to output features (shown in Fig. 3(a,b) as *Val OA-RV*). A similar analysis is performed for the arousal dimension. For the sequence at hand, in Fig. 3d we can see that 6 RVs in total are required for learning the arousal dimension. Note how for this prediction only one input feature RV is used. This implies that, only the actual input features (the facial expression features $\mathbf{x}$, in this case) from one frame (shown in Fig. 3d) are retained by the model.

The remaining 5 RVs centred around the output features are depicted in Fig. 3(a,b) as *Ar OA-RV*. An interesting observation is that, both for valence and arousal prediction, there are two common RVs centred around the output-features, in frame 1 and frame 15. In these frames, the arousal begins to decrease, and is accompanied by a change of sign in the valence dimension.

To conclude this section, in Fig. 3(e,f), we show the results of applying OA-RVM with a temporal window of $v = 4$ (Eq. 2). Note how the learned OA-RVM model provides a nearly perfect fit by using no more RVs than the original RVM model. Although the complexity of the model is observed to increase with an increase in the window size (Fig. 4 and Section VII-A), overall, the OA-RVM model appears to generalise to new data very well while avoiding overfitting.

VII. EXPERIMENTS AND RESULTS

In this section, we conduct both subject-dependent and subject-independent experiments to evaluate the proposed OA-RVM framework in terms of sparsity and prediction accuracy with respect to RVM and SVR.

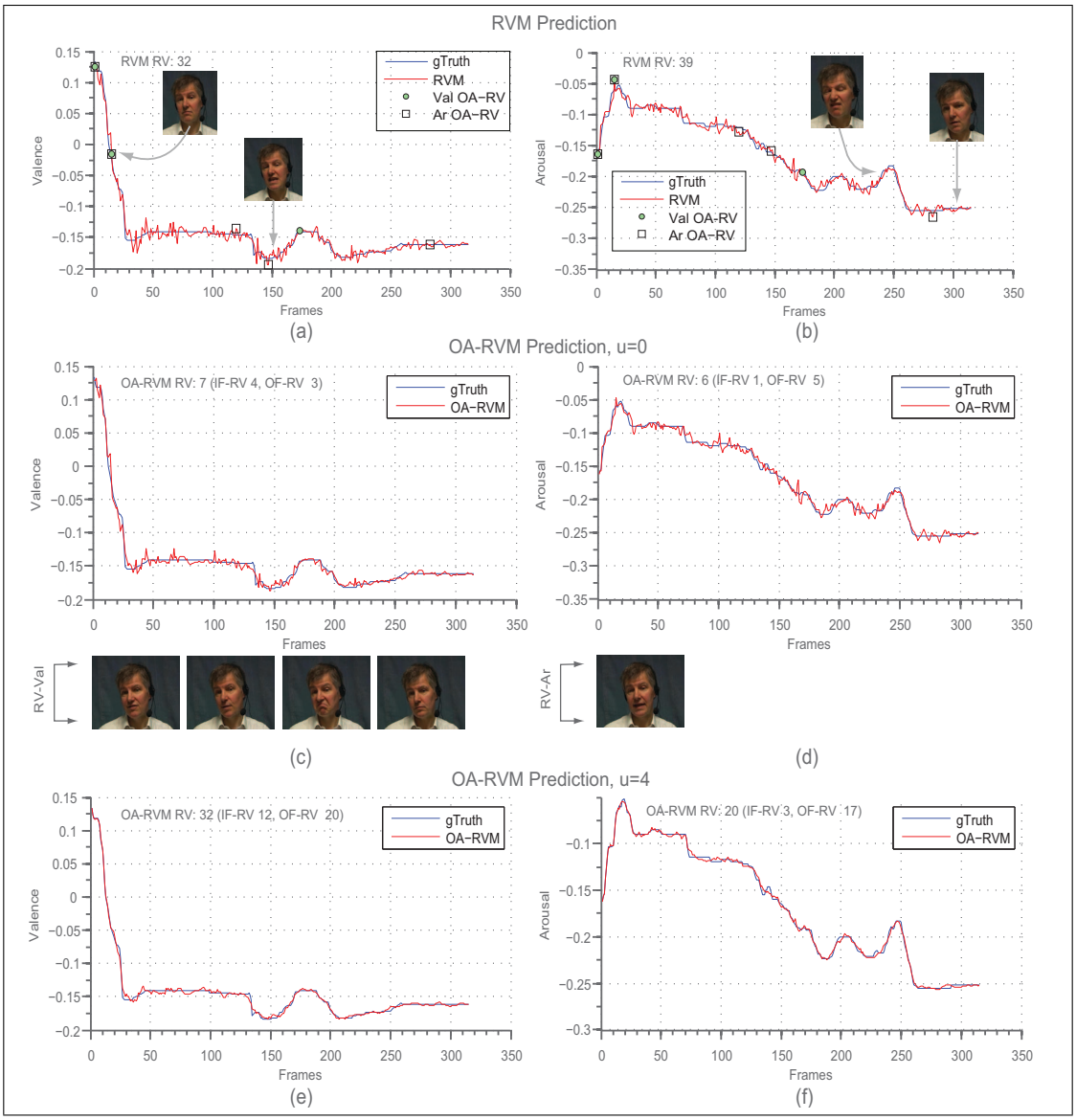

Fig. 3. Illustration of how employing the original RVM and the proposed OA-RVM provide continuous prediction of valence and arousal dimensions for one training sequence (315 frames). (a,b) RVM prediction with RVs used for OA-RVM, (c,d) OA-RVM prediction with a window of $v = 0$ and IF-RV frames, and (e,f) OA-RVM with prediction with a window of $v = 4$

A. *Sparsity*

This section provides a comparison between RVM and OA-RVM in terms of model sparsity. For this comparison, we use a small temporal window v, as a larger window complicates the model and increases the number of relevance vectors (RVs) needed. The comparison is performed by selecting the window with the highest sparsity while keeping the RMSE accuracy of both the RVM and the OA-RVM models approximately equal ($RMSE = 0.23$). The results are presented in Table I and Table II, and are discussed in Section VII-B.

Subject-dependent results are presented in Table I showing the number of relevance vectors selected by the traditional RVM and OA-RVM models. The most sparse results are achieved by using a window of $v = 0$. It can be clearly seen that both for valence and arousal, when employing the OA-RVM scheme, the number of RVs retained is decreased significantly.

The subject-independent results are presented in Table II. In this case, the results with highest sparsity were not always obtained by using a window of $v = 0$, but rather by using a window of $v = 1$ for subject 1, and $v = 2$ for subject 3 (positive class). An interesting observation is that more RVs are required for the negative class, leading to a more complex prediction model. Nevertheless, compared to the traditional RVM, the sparsity increase is still very high. Although we decomposed the RVs captured by the OA-RVM model into the ones that correspond to the input-features and to those corresponding to the output-features, we found no consistent patterns to report. Overall, we conclude that

subject-dependent variations in emotional expressions lead to variations in experimental results.

TABLE I

SUBJECT-DEPENDENT SPARSITY COMPARISON

	Valence$_{RV}$			Arousal$_{RV}$		
	RVM	OA-RVM	RMSE	RVM	OA-RVM	RMSE
Positive	267	10	0.23	270	12	0.22
Negative	245	10	0.23	244	13	0.36

TABLE II

SUBJECT-INDEPENDENT SPARSITY COMPARISON

	Valence$_{RV}$			Arousal$_{RV}$		
	RVM	OA-RVM	RMSE	RVM	OA-RVM	RMSE
Positive	485	10	0.2	495	11	0.15
Negative	394	21	0.19	417	29	0.36

B. Prediction

We begin our discussion on prediction accuracy of the proposed OA-RVM (with respect to RVM / SVR) by referring to the subject-dependent results presented in Table III.

For both valence and arousal dimensions, we observe that OA-RVM improves the prediction results in all cases. Arousal appears to be more challenging to model and predict for the negative class, in accordance with psychological evidence suggesting that visual cues are more indicative of valence rather than arousal [1]. Nevertheless, for the positive class, except for subject 4, arousal appears to be easier to model and predict.

The best prediction results are typically captured with an output-associative window size of $v > 8$, showing the significance of past and future context for continuous emotion prediction. To illustrate the increase of the RVs retained with the increase of window size, in Fig. 4 we present the number of RVs retained for subject 1 (positive class), from a window of $v = 0$ up to the optimal window of $v = 10$, which provided us with the best results. The increase in the number of RVs with the increasing window size applies to all subjects.

Overall, the optimal window size appears to be subject- and data-dependent. This in turn implies that naturalistic emotional displays are rather subject-specific in nature. For instance, predicting the valence and arousal level of subject 3, who displays the most subtle emotional expressions, appears to be easier compared to the rest of the subjects.

Table IV presents the subject-independent prediction results in terms of RMSE and window size (v). Each row on the table presents the results obtained by training the model using data from one subject (indicated in the first column) and using testing data from the rest.

OA-RVM provides better prediction results than RVM and SVR, in each and every tested case, similarly to the subject-dependent results. When comparing the RVM results to the results provided by SVR, it is possible to state that

TABLE III

SUBJECT-DEPENDENT PREDICTION RESULTS (RMSE).

	Valence			Arousal		
POS	**RVM**	**RVM-OA**	**v**	**RVM**	**RVM-OA**	**v**
subj1	0.16	0.15	10	0.13	0.11	10
subj2	0.17	0.13	18	0.14	0.13	5
subj3	0.11	0.09	12	0.10	0.09	18
subj4	0.17	0.15	8	0.23	0.19	18
NEG	**RVM**	**RVM-OA**	**v**	**RVM**	**RVM-OA**	**v**
subj1	0.14	0.10	12	0.30	0.29	14
subj2	0.11	0.09	18	0.37	0.33	9
subj3	0.08	0.07	18	0.22	0.21	18
subj4	0.11	0.10	18	0.48	0.40	12

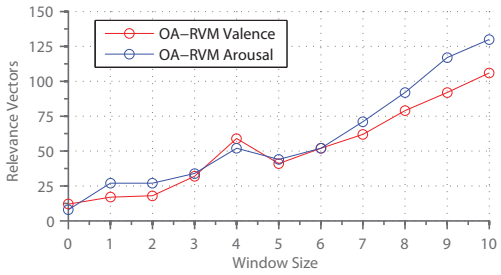

Fig. 4. Increase of Relevance Vectors in the OA model with the increase of window size in output features (Subject 1, Positive). The RVM RV are 643 and 626 for valence and arousal respectively.

on average, RVM performs better. However, there is no clear prediction advantage of one model over the other.

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, similarly to subject-dependent prediction results.

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. Exceptions can be observed in some experiments referring to subjects 3 and 4. Subject 4 who has the most intense expressions, is modelled with a smaller window for predicting positively valenced emotions and predicting arousal for negatively valenced emotions, while subject 3, who is observed to have the most subtle expressions is modelled with a smaller window for predicting negatively valenced emotions. Although these results are consistent, we do not consider them sufficient to draw general conclusions regarding the optimal window size with respect to the expressivity of each subject. We rather attribute them to subject and data-specific characteristics of the experiments.

Overall, naturalistic emotional expressions are highly subject-dependent [1]. However, from our experiments we conclude that automatic, subject-independent, dimensional and continuous prediction of emotions becomes feasible by utilising input and output associations as well as temporal context.

Psychological research findings suggest that there exist gender-related differences in expressing emotions (e.g.,

women appear to be more facially expressive than men [22]). However, in our experiments we found no consistent differentiations between male and female subjects.

To conclude this section, we comment on the noise aspect of the prediction (in terms of average standard deviation). For the subject-dependent experiments the average noise standard deviation for OA-RVM is 0.001, while for RVM is 0.007. For the subject-independent experiments the average noise standard deviation for OA-RVM is 0.003, while for RVM is 0.01. Thus, we are able to state that OA-RVM induces more confidence in the generated predictions than RVM.

.

TABLE IV

SUBJECT-INDEPENDENT PREDICTION RESULTS (RMSE)

	Valence				Arousal			
POS	**SVR**	**RVM**	**RVM-OA**	**v**	**SVR**	**RVM**	**RVM-OA**	**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**	**RVM-OA**	**v**	**SVR**	**RVM**	**RVM-OA**	**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

VIII. CONCLUSIONS AND DISCUSSION

In this paper, we proposed a novel Output-Associative Relevance Vector Machine (OA-RVM) regression framework that augments 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. We successfully applied the proposed framework for dimensional and continuous prediction of emotions from facial expressions, and demonstrated its advantages and efficiency over a comprehensive set of experiments, both for the commonly employed subject-dependent (training and testing the model for each subject separately) and the highly challenging subject-independent (training the model by using data from one subject only and testing on the rest) case. Our experimental 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. The size of the optimal temporal window may vary depending on the task and the data at hand.
- OA-RVM appears to provide a more sparse model than RVM, at no additional cost to the overall accuracy.
- Although there is an inherent, subject-dependent characteristic attributed to naturalistic emotional expressions; automatic, subject-independent, dimensional and continuous prediction of emotions is possible by utilising input and output associations, and temporal context.

As future work, the proposed model remains to be evaluated on databases with a larger number of subjects (e.g., SEMAINE) in order to (i) obtain deeper insights into the accuracy improvement provided by the OA-RVM model, and (ii) evaluate thoroughly the impact of the sparse OA-RVM model in terms of its generalisation capability over different data set(s) and subjects.

IX. ACKNOWLEDGMENTS

This work has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB) . The work of Hatice Gunes was funded by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE).

REFERENCES

[1] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[2] Z. Zeng et al., "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 39–58, 2009.

[3] J. A. Russell, "A circumplex model of affect," *Journal of Personality & Social Psychology*, vol. 39, pp. 1161–1178, 1980.

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

[5] A. Mehrabian and J. Russell, *An Approach to Environmental Psychology*. New York: Cambridge, 1974.

[6] A. M. Oliveira 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 the Int. Society for Psychophysics*, 2006, pp. 245–250.

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

[8] P. A. Lewis et al., "Neural correlates of processing valence and arousal in affective words," *Cerebral Cortex*, vol. 17, no. 3, pp. 742–748, Mar 2007.

[9] J. P. Weston et al., "Kernel dependency estimation," Germany, Tech. Rep. 98, August 2002.

[10] C. Cortes, M. Mohri, and J. Weston, "A general regression technique for learning transductions," in *Proc. of Int. conf. on Machine learning*. New York, NY, USA: ACM, 2005, pp. 153–160.

[11] L. Bo and C. Sminchisescu, "Structured output-associative regression," in *Proc. of CVPR*, 2009, pp. 2403–2410.

[12] ——, "Twin Gaussian Processes for Structured Prediction," *Int. Journal of Computer Vision*, vol. 87, pp. 28–52, 2010.

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

[14] G. Caridakis et al., "Modeling naturalistic affective states via facial and vocal expressions recognition," in *Proc. of ACM ICMI*, 2006, pp. 146–154.

[15] M. Wollmer, et al., "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. of Interspeech*, 2008, pp. 597–600.

[16] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion recognition space concept," *Proc. of European Signal Processing Conf.*, 2008.

[17] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[18] M. E. Tipping and A. Faul, "Fast marginal likelihood maximisation for sparse bayesian models," in *Proc. of Int. Workshop on Artificial Intelligence and Statistics*, 2003, pp. 3–6.

[19] E. Douglas-Cowie et al., "The humaine database: addressing the needs of the affective computing community," in *Proc. of Int. Conf. on Affective Computing & Intelligent Interaction*, 2007, pp. 488–500.

[20] 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: Advances in Capturing, Coding and Analyzing Multimodality*, 2010, pp. 43–48.

[21] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features," in *Proc. of IEEE Int. Conf. on Automatic Face & Gesture Recognition*, 2004, pp. 97–102.

[22] A. M. Kring and A. H. Gordon, "Sex differences in emotion: Expression, experience, and physiology,," *Journal of Personality & Social Psychology*, vol. 74, no. 3, pp. 686 – 703, 1998.