Image and Vision Computing xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Output-associative RVM regression for dimensional and continuous emotion prediction ☆

Mihalis A. Nicolaou [a,*], Hatice Gunes [b], Maja Pantic [a,c]

[a] Department of Computing, Imperial College London, UK
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[c] EEMCS, University of Twente, The Netherlands

ARTICLE INFO

Article history:
Received 27 July 2011
Received in revised form 8 December 2011
Accepted 16 December 2011
Available online xxxx

Keywords:
Dimensional and continuous emotion prediction
Facial expressions
Shoulder movements
Audio cues
Output-associative RVM regression

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 spatial dependencies* between the output vectors, as well as *repeating output patterns* and *input–output associations*, that can provide more robust and accurate predictors when modeled 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 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, and evaluate the proposed framework by focusing on the case of multiple nonverbal cues, namely facial expressions, shoulder movements and audio cues. We demonstrate the advantages of the proposed OA-RVM regression by performing subject-independent evaluation using the SAL database that constitutes naturalistic conversational interactions. The experimental results show that OA-RVM regression outperforms the traditional RVM and SVM regression approaches in terms of accuracy of the prediction (evaluated using the Root Mean Squared Error) and structure of the prediction (evaluated using the correlation coefficient), generating more accurate and robust prediction models.

1. 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 normalization, 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 a variance-and-bias-based evaluation with Root Mean Squared Error (RMSE, i.e., how much prediction and ground-truth values vary), but also with a structure-based evaluation with the correlation coefficient (COR, i.e., evaluating the covariance of the prediction with the ground truth), resulting in a more accurate and robust model. In order to evaluate whether the proposed technique's performance is *cue and modality invariant*, we focus on a highly challenging, yet a very suitable problem: dimensional and continuous emotion prediction from nonverbal *heterogeneous* cues (i.e., facial expressions, shoulder movements and audio cues).

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

☆ This paper has been recommended for acceptance by Jan-Michael Frahm.

* Corresponding author. Tel.: +44 72075941117.
E-mail address: mihalis@imperial.ac.uk (M.A. Nicolaou).

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2011.12.005

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 active or inactive 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 quantized recognition) and output-associative structured prediction are two highly inter-related problems. Psychological evidence has shown that the V-A dimensions are inter-correlated [6,7,4,8]. Therefore, the proposed framework 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).

2. Related work

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.

2.1. Output-associative structured regression

Output-associative structured regression has gained much popularity 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 optimization 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] optimizes 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 employs 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 minimization of the KL divergence. Both works have been applied to modeling human pose estimation.

We choose to extend RVM as it is considered to be more efficient than traditional GP, and is known to provide a sparse solution. Note that other works on extending RVM have also been proposed, e.g. [13] proposed a robust RVM which models outlier noise while [14] proposed a multi-variate version of RVM.

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 provides explicit modeling of the noise component.

2.2. Dimensional and continuous emotion prediction

Dimensional and continuous recognition or prediction of emotions is a relatively unexplored area in the field of affective computing, with the first workshop organized on this topic only recently [15].

To date, the most commonly employed strategy in automatic dimensional affect recognition from visual signals has been to reduce the recognition problem to a two-class problem (positive vs. negative or active vs. passive classification; e.g., [16,17]) or a four-class problem (classification into the quadrants of 2D A-V space; e.g., [18,19]).

Currently, there are also a number of works focusing on dimensional and continuous prediction of emotions from the visual modality [20,21]. The work by Gunes and Pantic focuses on dimensional prediction of emotions from spontaneous conversational head gestures by mapping the amount and direction of head motion, and occurrences of head nods and shakes into arousal, expectation, intensity, power and valence level of the observed subject using SVRs [20].

Similarly to the affect recognition from visual signals, the most commonly employed strategy in automatic dimensional affect recognition from audio signals has been to reduce the recognition problem to a two-class problem (positive vs. negative or active vs. passive classification; e.g., [22]) or a four-class problem (classification into the quadrants of 2D arousal-valence (A-V) space; e.g., [23]).

As far as actual continuous dimensional affect prediction (without quantization) is concerned, there exist a number of methods that deal exclusively with speech (i.e., [23–25]). The work by Wöllmer et al. uses the SAL Database and Long Short-Term Memory neural networks and Support Vector Machines for Regression (SVR) [24]. Grimm and Kroschel use the Vera am Mittag database [26] and SVRs, and compare their performance to that of the distance-based fuzzy k-Nearest Neighbor and rule-based fuzzy-logic estimators [25]. The work by Espinosa et al. also use the Vera am Mittag database [26] and examine the importance of different groups of speech acoustic features in the estimation of continuous PAD dimensions (Pleasure, Arousal and Dominance) [27].

When it comes to dimensional emotion recognition using multiple modalities the focus has mainly been on discriminating between more coarse categories, such as positive vs. negative [16] and active vs. passive [28]. Of these, Caridakis et al. [28] use the SAL database, combining auditive and visual modalities. Nicolaou et al. focus on audio-visual classification of spontaneous affect into negative or positive emotion categories using facial expression, shoulder and audio cues, and utilizing 2- and 3-chain coupled Hidden Markov Models and likelihood space classification to fuse multiple cues and modalities [16]. Kanluan et al. [29] combine audio and visual cues for affect recognition in A-V space by fusing facial expression and audio cues, using SVRs and late fusion with a weighted linear combination with discretized labels (on a 5-point scale in the range of $[-1,+1]$ for each emotion dimension). Wöllmer et al. use multimodal acted data that contain face (obtained from motion capture and video) and audio information, and recognize 3–5 levels of A-V values using various classification techniques [30]. More recent works focus on dimensional and continuous prediction of emotions from multiple modalities. For instance, Eyben et al. [31] propose a string-based approach for fusing the behavioral events from visual and auditive modalities (i.e., facial action units, head nods and shakes, and verbal and nonverbal audio cues) to predict human affect in a continuous dimensional space (in terms of arousal, expectation, intensity, power and valence dimensions). Metallinou et al. in [32] focus on analyzing the vocal and body language behavior (via MoCap features) of pairs of actors improvising diadic interactions. For each actor's recording, they computed the Spearman correlation coefficient between the mean annotation and the MLE curve. Activation and dominance were predicted from visual and audio-visual cues reasonably well. However, for valence, the MLE mapping curves failed to track the changes and the respective median correlations were close to zero. Another representative approach is that of

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005

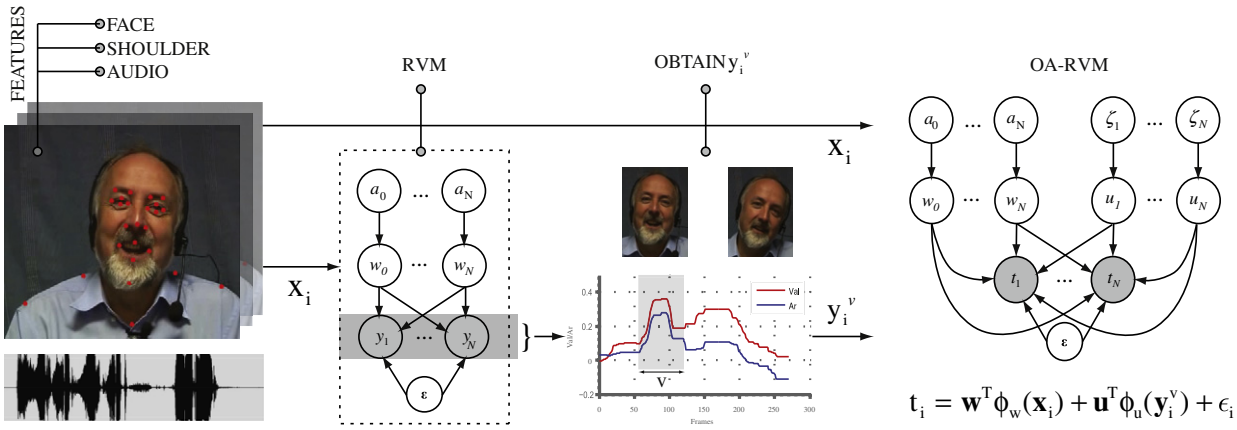

Fig. 1. Outline of the proposed method. The tracked features (from facial expressions, shoulder movements and audio) are fed into an initial regressor (here, RVM) to obtain an initial prediction. A temporal window v is applied on the multi-dimensional output of valence and arousal, constructing the output feature vectors ($\mathbf{y_i^v}$). Both the input features $\mathbf{x_i}$ and the output features $\mathbf{y_i^v}$ are fed into the OA-RVM model which provides the final prediction.

Gilroy et al. [33] that propose a dimensional multimodal fusion scheme based on the PAD space to support detection and integration of spontaneous affective behavior of users (in terms of audio, video and attention events) experiencing arts and entertainment.

Despite the increased interest in dimensional representation of emotions, none of the works proposed to date have explored input–output associations and spatio–temporal dependencies between the output vectors for dimensional and continuous emotion prediction.

2.3. Contributions

This work has been inspired by the pioneering works of [34] and [35] that capitalized on the fact that the arousal and valence dimensions are correlated, and presented an approach that fuses spontaneous facial expression, shoulder movement and audio cues for dimensional and continuous prediction of emotions in valence-arousal space. They proposed an output-associative fusion framework that incorporates correlations between emotion dimensions. Their findings suggested that incorporating correlations between affect dimensions provides greater accuracy for continuous affect prediction.

Building upon the idea that arousal and valence dimensions are correlated, we further explore input–output associations and spatio–temporal dependencies between the output vectors. 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 heterogeneous nonverbal cues.

An earlier version of this paper appeared in the Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition [21]. There are two major deviations from the previous work that merit being highlighted: 1) the previous work did not investigate the feasibility and the usefulness of the proposed OA-RVM framework on dimensional and continuous prediction of emotions from nonverbal *heterogeneous* cues (the focus was only on facial expression cues); and 2) the proposed model was not evaluated in terms of the prediction covariance with respect to the ground truth, a measure of structural agreement of the two signals. The current paper addresses these issues by exploring how the behavior of the OA-RVM model changes (in terms of prediction accuracy and spatio–temporal structure) depending on the expressive cue/modality employed, by adopting a leave-one-subject-out (subject-independent) experimental setup.

The outline of the proposed method is presented in Fig. 1. The tracked/extracted features (from facial expressions, shoulder movements and audio) are fed into an initial (cue-specific) regressor, which in our case is chosen to be RVM (trained separately for each cue). An initial, noisy prediction is obtained by RVM. A temporal window v is applied on the multi-dimensional output of valence and arousal, thus constructing a set of new vectors which we call output features ($\mathbf{y_i^v}$). Both the input features $\mathbf{x_i}$ and the output features $\mathbf{y_i^v}$ are fed into the OA-RVM model which learns specific weights for each input and output feature vector. The final prediction is a linear combination of the kernel-projected input and output features.

The rest of the paper is organized as follows. In Section 3, we briefly revisit the RVM and SVM models in order to provide a basis for OA-RVM, introduced and explained in Section 4. Section 5 describes the data set employed in our experiments, as well as the feature extraction and tracking process. Section 6 provides a demonstration of the behavior of the model when learning continuous emotion dimension values, while Section 7 presents the experiments and discusses the results. Finally, Section 8 concludes the paper.

3. 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, ($\mathbf{x_i}$, $\mathbf{t_i}$).[1] In the Bayesian framework applied in RVM, our goal is to learn the functional

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

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

The graphical model of RVM is presented in Fig. 2.

[1] We denote that $\mathbf{t_i}$ is a multidimensional vector containing all the values to be predicted for each frame (in our case, both valence and arousal). Nevertheless, the methods we apply are inherently single output methods. Thus, a different function is learnt for each output dimension (t_i).

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005

In SVR, the functional

$$t_i = \mathbf{w}^T \phi(\mathbf{x}_i) + b$$

is learned, where ϕ is an implicit mapping to a kernel space, $\mathbf{w}$ represents the set of weights, and b represents the bias. Lagrangian optimization is employed to determine the optimal parameters of the problem. In contrast to Bayesian regression methods, there is no explicit noise modeling in SVR. The structural risk minimization principle is applied to minimize the risk of overfitting.

4. Output-associative RVM regression

In this section we describe the proposed OA-RVM framework. The graphical models of both OA-RVM and RVM are illustrated in Fig. 2, while an overview of the algorithm for training and prediction in OA-RVM framework is presented in Alg. 1.

Firstly, to obtain the output associative functional, we increment Eq. (1) as follows:

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

Where each $\mathbf{y_i^v}$ is a vector of multi-dimensional outputs over a temporal window of $[i-v, i+v]$[2] 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).[3]

4.1. The framework

In this section we describe the Bayesian framework that our model is based on. Firstly, we consider $\Phi_{\mathbf{w}}$ ($N \times M_w$) to be the basis matrix attained by applying a selected kernel to the input features $\mathbf{x}$, and $\Phi_{\mathbf{u}}^{\mathbf{v}}$ (Φ_u^v ($N \times M_u$)) respectively, for the output features $\mathbf{y^v}$ (M_u and M_w, referring to the number of basis vectors). Then, by extending Eq. (2) we obtain:

$$\mathbf{t} = \Phi_{\mathbf{w}}\mathbf{w} + \Phi_{\mathbf{u}}^{\mathbf{v}}\mathbf{u} + \varepsilon = \Phi_{\mathbf{wu}}\mathbf{w_u} + \varepsilon \tag{3}$$

where $\Phi_{\mathbf{wu}} = [\Phi_{\mathbf{w}} | \Phi_{\mathbf{u}}^{\mathbf{v}}]$ is the $N \times (M_w + M_u)$ OA-RVM design matrix:

$$\Phi_{\mathbf{wu}} = \begin{bmatrix} K_w(\mathbf{x}_1, \mathbf{x}_1) & \dots & K_w(\mathbf{x}_1, \mathbf{x_n}) & K_u(\mathbf{y}_1^{\mathbf{v}}, \mathbf{y}_1^{\mathbf{v}}) & \dots & K_u(\mathbf{y}_1^{\mathbf{v}}, \mathbf{y_n^v}) \\ \vdots & & \vdots & \vdots & & \vdots \\ K_w(\mathbf{x_n}, \mathbf{x}_1) & \dots & K_w(\mathbf{x_n}, \mathbf{x_n}) & K_u(\mathbf{y_n^v}, \mathbf{y}_1^{\mathbf{v}}) & \dots & K_u(\mathbf{y_n^v}, \mathbf{y_n^v}) \end{bmatrix}$$

with K_w and K_u being the kernel applied to input and output features respectively. Typically, an extra unit column is appended to the kernel to account for the bias. Furthermore, $\mathbf{w_u} = \left[\mathbf{w}_1 \dots \mathbf{w}_{\mathbf{M_w}} \middle| \mathbf{u}_1 \dots \mathbf{u}_{\mathbf{M_u}}\right]^T$ represents the concatenated vector of weights. Thus, the complete data set likelihood is formulated as:

$$\begin{aligned} P\left(\mathbf{t} \middle| \mathbf{w}, \mathbf{u}, \sigma^2\right) &= \prod_{i=1}^{N} N\left(\mathbf{w}^T \phi_w(\mathbf{x_i}) + \mathbf{u}^T \phi_u(\mathbf{y_i^v}), \sigma^2\right) \\ &= \prod_{i=1}^{N} N\left(\mathbf{w_u}^T [\phi_w(\mathbf{x_i}) | \phi_u(\mathbf{y_i^v})], \sigma^2\right) \end{aligned} \tag{4}$$

Following the Bayesian approach of RVM [36], 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:

$$P(\mathbf{w}|\alpha) = \prod_{i=0}^{M_u} N\left(0, \alpha_i^{-1}\right) \tag{5}$$

$$P(\mathbf{u}|\zeta) = \prod_{i=1}^{M_w} N\left(0, \zeta_i^{-1}\right) \tag{6}$$

We have now introduced two vectors of hyperparameters, α controlling the distribution of the weights $\mathbf{w}$ (as originally used in RVM), and ζ controlling the distribution of the weights $\mathbf{u}$ (for our output features).

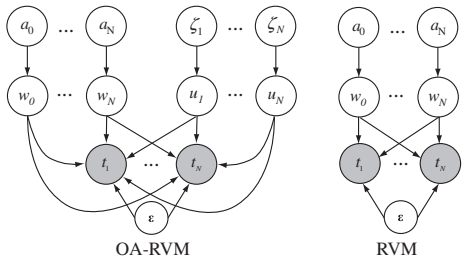

Fig. 2. Graphical model comparison of RVM and OA-RVM. Shaded nodes are observed variables.

4.2. Inference

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

$$P\left(\mathbf{w}, \mathbf{u}, \alpha, \zeta, \sigma^2 | \mathbf{t}\right) = \frac{P\left(\mathbf{t} | \mathbf{w}, \mathbf{u}, \alpha, \zeta, \sigma^2\right) P\left(\mathbf{w}, \mathbf{u}, \alpha, \zeta, \sigma^2\right)}{p(\mathbf{t})} \tag{7}$$

Ideally, given a new test data x, we would like to predict target t:

$$p(t_* | \mathbf{t}) = \int P\left(\mathbf{t}_* | \mathbf{w}, \mathbf{u}, \alpha, \zeta, \sigma^2\right) P\left(\mathbf{w}, \mathbf{u}, \alpha, \zeta, \sigma^2 | \mathbf{t}\right) d\mathbf{w} d\mathbf{u} d\alpha d\zeta d\sigma^2 \tag{8}$$

Unfortunately, the above equation is intractable, thus an approximation is needed. Therefore, similarly to the original RVM formulation [36], we decompose the posterior as follows:

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

Using the Bayes theorem we obtain:

$$P\left(\mathbf{w}, \mathbf{u} | \mathbf{t}, \alpha, \zeta, \sigma^2\right) = \frac{P\left(\mathbf{t} | \mathbf{w}, \mathbf{u}, \sigma^2\right) P(\mathbf{w}, \mathbf{u} | \alpha, \zeta)}{P\left(\mathbf{t} | \alpha, \zeta, \sigma^2\right)} \tag{10}$$

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 $P(\mathbf{w}, \mathbf{u} | \alpha, \zeta)$, a zero-mean Gaussian distribution with a covariance matrix $\mathbf{A_Z} = \mathrm{diag}(\alpha_1 \dots \alpha_{M_w}, \zeta_1 \dots \zeta_{M_u})$.

$$P\left(\mathbf{t} | \alpha, \zeta, \sigma^2\right) = \int P\left(\mathbf{t} | \mathbf{w}, \mathbf{u}, \sigma^2\right) P(\mathbf{w}, \mathbf{u} | \alpha, \zeta) d\mathbf{w} d\mathbf{u} \tag{11}$$

is a convolution of Gaussian and after replacing with $\mathbf{w_u}$, $\mathbf{A_Z}$ and $\Phi_{\mathbf{wu}}$, it can be shown [36] 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. (10) is considered to be a Gaussian distribution with a mean $\mu = \sigma^2 \Sigma \Phi_{wu}^T \mathbf{t}$ and a covariance matrix $\Sigma = \left(\mathbf{A_Z} + \sigma^2 \Phi_{wu}^T \Phi_{wu}\right)^{-1}$.

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

[3] 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. $\varepsilon_i \sim N(0, \sigma_y^2 + \sigma_x^2) = N(0, \sigma^2)$.

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005

Returning to the second component $P(\alpha,\zeta,\sigma^2|\mathbf{t})$ of the posterior in Eq. (9), by following the Bayes rule, we find it to be proportional to:

$$P\left(\alpha,\zeta,\sigma^2|\mathbf{t}\right)\propto P\left(\mathbf{t}|\alpha,\zeta,\sigma^2\right)P(\alpha)P(\zeta)P\left(\sigma^2\right) \tag{12}$$

By assuming uniform uninformative hyperpriors [36], we need to maximize the marginal likelihood, $P(\mathbf{t}|\alpha,\zeta,\sigma^2)$ with respect to the hyperparameters. Again, we have a convolution of Gaussians (Eq. (11)) which in turn generates another zero mean Gaussian distribution with covariance matrix $\sigma^2\mathbf{I}+\Phi_{wu}\mathbf{A_z}^{-1}\Phi_{wu}^T$. The maximization of the marginal likelihood can be performed by expectation maximization as described in [36] or the faster marginal maximization algorithm proposed in [37]. The most probable values (*MP*) are selected by the chosen optimization procedure ([36,37]), while we adopt an approximation of $P(\alpha,\zeta,\sigma^2|\mathbf{t})$ in Eq. (9) by replacing the distribution with a delta function at its mode.

4.3. 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 $\alpha_{\mathbf{z}}=[a_1\ldots a_{M_w},\zeta_1\ldots\zeta_{M_u}]$ and using Eqs. (8) and (10) we obtain:

$$P\left(t_*|\mathbf{t},\alpha_{\mathbf{z}MP},\sigma^2_{MP}\right)=\int P\left(t_*|\mathbf{w_u},\sigma^2_{MP}\right)P\left(\mathbf{w_u}|\mathbf{t},\alpha_{\mathbf{z}MP},\sigma^2_{MP}\right)d\mathbf{w_u} \tag{13}$$

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

$$P\left(t_*|\mathbf{t},\alpha_{\mathbf{z}MP},\sigma^2_{MP}\right)\sim N\left(t_*|\sigma^2_*\right) \tag{14}$$

where

$$t_*=\mu_{wu}^T[\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^{\mathbf{v}})] \tag{15}$$

$$\sigma^2_*=\sigma^2_{MP}+[\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^{\mathbf{v}})]^T\Sigma[\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^{\mathbf{v}})] \tag{16}$$

with t being the test point prediction, and σ^2 being the prediction variance (relating to *confidence* in the obtained prediction). The parameter vector μ_{wu} contains the weights for the input and output relevance vectors, i.e. $\mu_{wu}=[\mu_w|\mu_u]$. 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 model of OA-RVM with respect to the original RVM can be seen in Fig. 2. An overview of the OA-RVM training and prediction procedures is presented in Algorithm 1.

Algorithm 1. OA-RVM algorithm

Training. Data: $(\mathbf{x_i},\mathbf{t_i})$, i = 1,…, N

1. Obtain output features $\mathbf{y_i^v}$
2. Construct basis matrix $\Phi_{\mathbf{wu}}=[\Phi_{\mathbf{w}}|\Phi_{\mathbf{u}}^{\mathbf{v}}]$
   - **2a.** Apply kernel K_w for obtaining $\Phi_{\mathbf{w}}$ for input features $\mathbf{x}$
   - **2b.** Apply kernel K_u for obtaining $\Phi_{\mathbf{u}}^{\mathbf{v}}$ for output features $\mathbf{y^v}$
3. Marginal likelihood maximization
   - **3a.** Determine hyperparameters (α,ζ,σ^2)
   - **3b.** $\mu=\sigma^2\Sigma\Phi_{wu}^T\mathbf{t}$, $\Sigma=\left(\mathbf{A_Z}+\sigma^2\Phi_{wu}^T\Phi_{wu}\right)^{-1}$

Prediction for test point $\mathbf{x}_*$:

1. Obtain output features $\mathbf{y}_*^{\mathbf{v}}$
2. Predict and estimate variance:
   - **2a.** $t_*=\mu_{wu}^T[\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^{\mathbf{v}})]$
   - **2b.** $\sigma^2_*=\sigma^2_{MP}+[\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^{\mathbf{v}})]^T\Sigma[\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^{\mathbf{v}})]$

4.4. Window size

The output feature window length v for OA-RVM is treated as a regular parameter in the framework. Therefore, many heuristics and validation techniques can be employed in order to define the parameter for a given training set. The most direct method would be to perform cross-validation (i.e. similarly to SVM) in order to determine the optimal value for the specific error metric employed. Another way is to compare the maximized marginal likelihood of each model trained with a specific window size (i.e. a maximum likelihood test). Assuming we have a set V of windows to be evaluated, for each $v_i\in V$ the marginal likelihood $L_{v_i}\sim N(0,\sigma^2\mathbf{I}+[\Phi_{\mathbf{w}}|\Phi_{\mathbf{u}}^{\mathbf{v_i}}]\mathbf{A_z}^{-1}[\Phi_{\mathbf{w}}|\Phi_{\mathbf{u}}^{\mathbf{v_i}}]^T)$ is maximized. The window size providing the maximum likelihood can then be selected, i.e. $v=argmax_{v_i}L_{v_i}$.

4.5. A generalized view

In this section we aim to provide a more general perspective of the proposed framework while comparing it to other static regression frameworks (e.g. SVM and RVM).

In a typical static regression framework (e.g. SVM and RVM), we consider only the current input to participate in the prediction, i.e.

$$P(t_i|\mathbf{x_1}\ldots\mathbf{x_i}\ldots\mathbf{x_N})=P(t_i|\mathbf{x_i})$$

In the proposed framework, each prediction not only depends on the current input but also on the output features, which essentially represent a *temporal* noisy version of the targets to be estimated:

$$P(t_i|\mathbf{x_1}\ldots\mathbf{x_i}\ldots\mathbf{x_N})=P(t_i|\mathbf{x_i},\mathbf{y_i^v})$$

The output features $\mathbf{y_i^v}$ represent a noisy prediction of the targets over time (a pre-defined temporal window). Therefore,

$$P(t_i|\mathbf{x_1}\ldots\mathbf{x_i}\ldots\mathbf{x_N})=P(t_i|\mathbf{x_i},\hat{\mathbf{t}}_{i-v},\ldots,\hat{\mathbf{t}}_i,\ldots,\hat{\mathbf{t}}_{i+v})$$

where each $\hat{\mathbf{t}}_i$ is the noisy prediction of $\mathbf{t_i}$ at input datum i. The prediction is thus conditioned both on the current input frame, as well as the noisy prediction of the multi-dimensional targets over the specified temporal window.

Conditioning on the intermediate noisy predictions can be considered as a form of *ensemble learning*, specifically of *stacked generalization* [38,39] with continuous labels. A specific stacked generalization algorithm could also be investigated for training OA-RVM to obtain insight on its benefits for method generalization.

4.6. Complexity

The optimization algorithm of RVM generally involves the optimization of a non-convex function. The inversion of an *MxM* matrix is required, where M is the number of basis vectors in the model, thus inducing $O(M^3)$ computational complexity. In OA-RVM, without loss of generality, we assume that we have *2M* basis vectors: 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, the output features in OA-RVM are obtained by utilizing 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)*. In conclusion, the theoretical complexity of OA-RVM is of the same order as RVM. Nevertheless, in practice OA-RVM has a higher computational complexity than RVM, since it involves executing the original RVM algorithm as well as OA-RVM, which implies an augmented kernel with twice the number of candidate basis vectors compared to RVM.

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005

5. Data set and feature extraction

As a proof of concept, the proposed OA-RVM regression is applied to the highly challenging problem of *dimensional and continuous prediction* of emotions from heterogeneous nonverbal cues, namely facial expressions, shoulder movements and audio-cues. Our aim is to explore how the behavior of the OA-RVM model changes (in terms of prediction accuracy and spatio–temporal structure) depending on the expressive cue/modality employed.

5.1. Data set

For experimental validation we use the Sensitive Artificial Listener (SAL) Database [40]. 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 take advantage only 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).

Example frames from this portion of the SAL database, together with the trackings of facial points, are shown in Fig. 3. Based on the annotations provided, we used a set of automatic segmentation and ground truth generation algorithms [41] 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.

5.2. Facial expressions

For extracting facial expression features, we employ the Patras–Pantic particle filtering tracking scheme [42] for tracking the facial feature movements displayed during the naturalistic interactions. We track the corners of the eyebrows (4 points), the eyes (8 points), the nose (3 points), the mouth (4 points) and the chin (1 point). For each video segment containing n frames, the tracker results in a feature set with dimensions $n*20*2$. 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). Fig. 3(a) shows examples from the data set employed together with the tracking of the facial feature points.

5.3. Shoulder movements

The motion of the shoulders is captured by tracking 2 points on each shoulder and one stable point on the torso, usually just below the neck (see Fig. 3(b)). We initialize the tracked points in the first frame of each sequence manually, while the standard Auxiliary Particle Filtering (APF) [43] is subsequently used to track the shoulder points. This scheme is less complex and faster compared to the Patras–Pantic particle filtering tracking scheme, it does not require learning the model of prior probabilities of the relative positions of the shoulder points, while resulting in sufficiently high accuracy. For each video segment containing n frames, the tracker results in a feature set with dimensions $n*5*2$. The SAL database consists of challenging data with sudden body movements and out-of-plane head rotations. As the focus of this paper is on dimensional and continuous affect prediction, we would like to minimize the effect of imperfect and noisy point tracking on the automatic prediction. Therefore, both facial point tracking and shoulder point tracking have been done in a semi-automatic manner (with manual correction when tracking is imperfect).

5.4. Audio features

Our audio features include Mel-frequency Cepstrum Coefficients (MFCC, MFCC-Delta) [44] and prosody features (the energy of the signal, the root mean squared energy and the pitch obtained by using a Praat pitch estimator [45]).

We used 6 cepstrum coefficients, thus obtaining 6 MFCC and 6 MFCC-Delta features for each audio frame. We have essentially extracted the typical set of features used by other works (e.g., [46]) for automatic affect recognition. Along with pitch, energy and RMS energy, we obtained a set of features with dimensionality 15 (per audio frame).

6. 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. 4. The figure illustrates how employing the original RVM and the proposed OA-RVM

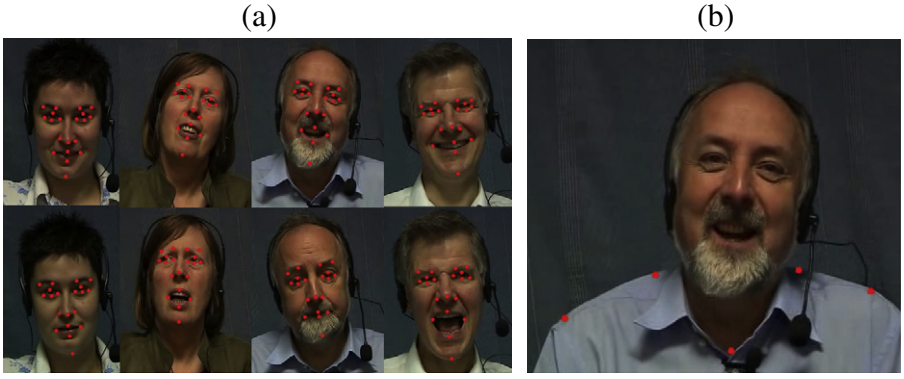

Fig. 3. Examples of the data at hand from the SAL database. (a) Facial expression tracking (20 points) (b) Shoulder tracking (5 points).

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005

provides continuous prediction of valence and arousal dimensions for one training sequence (consisting of 315 frames) extracted as explained in Section 5.

The predictions generated by RVM are shown in Fig. 4(a,b) while the OA-RVM generated predictions with a window of $v=0$ and $v=4$ are shown in Fig. 4(c,d) and (e,f), respectively. The ground truth for both the valence and the arousal dimensions is shown in all figures as *gTruth*, for comparison purposes. The generated predictions for valence appear on the left column of Fig. 4, 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. 4a) and 39 RVs for arousal (Fig. 4b). Fig. 4(c,d)

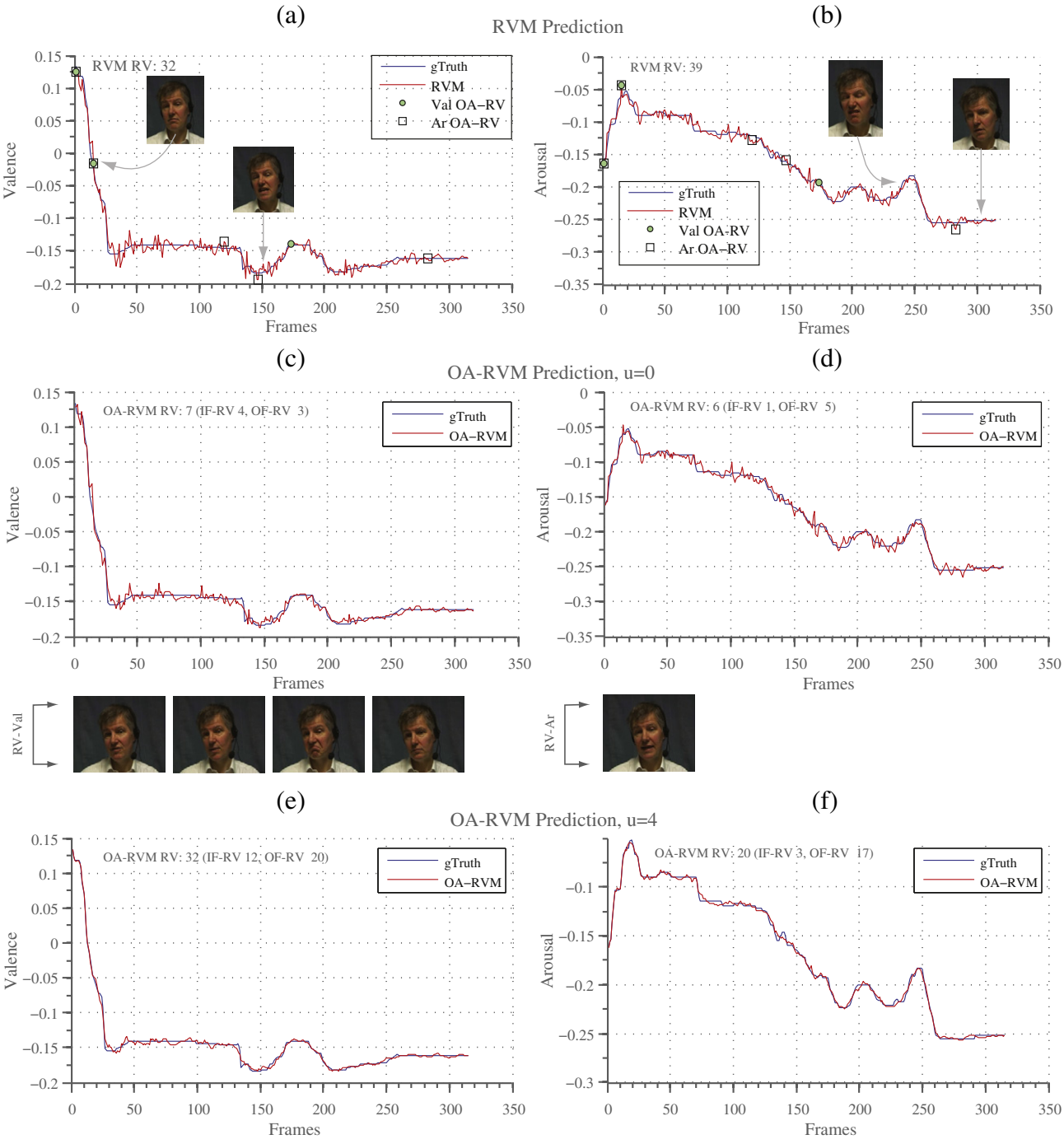

Fig. 4. Illustration of how employing the original RVM and the proposed OA-RVM provide continuous prediction of valence and arousal dimensions for one training sequence (315 frames). (a,b) RVM prediction with RVs used for OA-RVM, (c,d) OA-RVM prediction with a window of v = 0 and IF-RV frames, and (e,f)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

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005

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 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 ($\mathbf{y}^{\mathbf{v}}$, depicted as *OF* in the figure). To illustrate the behavior of the framework, we decompose the relevance vectors (RVs) selected by OA-RVM into the RVs centered around the input features (RV-IF) and the RVs centered 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. 4c) and 3 RVs corresponding to output features (shown in Fig. 4(a,b) as *Val OA-RV*). A similar analysis is performed for the arousal dimension. For the sequence at hand, in Fig. 4d we can see that 6 RVs are needed by OA-RVM. Note how for this prediction, only one input feature RVis used by OA-RVM. The remaining 5 RVs centered around the output features are depicted in Fig. 4(a,b) as *Ar OA-RV*. An interesting observation is that, in frame 1 and frame 15 the arousal begins to decrease, and is accompanied by a change of sign in the valence dimension. The OA-RVM framework is able to capture this in its valence and arousal prediction via two common RVs centered around the output-features in frame 1 and frame 15.

To conclude this section, in Fig. 4(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 (from which the output features are generated). Yet, both the MSE and COR metrics are improved. Although the complexity of the model is observed to increase with an increase in the window size, overall, the OA-RVM model appears to generalize to new data very well while avoiding overfitting.

7. Experiments and results

7.1. Experimental setting

We apply the proposed OA-RVM regression to the highly challenging problem of *dimensional and continuous prediction* of emotions from heterogeneous nonverbal cues, namely facial expressions, shoulder movements and audio cues. Our aim is to conduct comprehensive experiments in order to explore how the behavior of the OA-RVM model changes (in terms of prediction accuracy and spatio–temporal structure) depending on the expressive cue/modality employed.

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, $K(x,x_i)=exp\left\{\left(-(\mathbf{x}-\mathbf{x_i})^2\right)/r^2\right\}$ 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 $[0-30]$ and 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 fast marginal likelihood maximization algorithm proposed in [37]). We use RVM to obtain the initial (noisy) 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 5) have been coarsely classified into either positive or negative, prior to the prediction (regression) procedure. [4]The classification stage is beyond the scope of this paper, and can be achieved by applying an accurate (coarse) classifier, e.g. [16], as the basis for the current framework. This assumption is motivated by the fact that we would like to focus on the prediction results in 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 experiments using *subject-independent* cross-validation, where we train the model using data from three subjects and evaluate the trained model using the data from the subject left out for evaluation. Results are averaged across four-fold subject-independent cross-validation.

Note that subject-independent evaluation using this database is considered highly challenging [24] as annotated data is only available for a small number of subjects. More specifically, during training, the model is able to learn only a limited (subject-specific) subspace of the human affective variability. Moreover, performing regression in a continuous space (rather than classification into a predetermined set of labels) poses additional challenges.

As evaluation metrics we use both the root mean squared error (RMSE) and the correlation (COR) between the prediction and the ground truth values. RMSE evaluates the prediction by taking into account the squared error of the prediction from the ground truth. As discussed in [34], the RMSE, which represents the bias error and variance of the prediction, can be misleading with regard to how realistic the prediction of a regression technique can be. The correlation coefficient (COR) provides an evaluation of the linear relationship between the prediction and the ground truth, and subsequently, an evaluation of whether the model has managed to capture the linear structural patterns inhibited in the data at hand.

Table 1
Evaluating SVM, RVM and OA-RVM in terms of RMSE for predicting arousal and valence from face, shoulder and audio cues. Results are averaged across four-fold subject-independent cross-validation. Best results are indicated in bold.

		Valence			Arousal		
Class	Cue	SVM	RVM	OA-RVM	SVM	RVM	OA-RVM
POS	Face	0.200	0.166	**0.160**	0.157	0.166	**0.147**
	Shoulders	0.257	0.177	**0.171**	0.164	0.146	**0.132**
	Audio	0.176	0.179	**0.171**	0.146	0.144	**0.130**
NEG	Face	0.150	0.940	**0.088**	0.365	0.374	**0.342**
	Shoulders	0.110	0.103	**0.097**	0.355	0.371	**0.354**
	Audio	0.101	0.102	**0.097**	0.339	0.339	**0.300**

7.2. Experimental results and analysis

In this section, we will discuss the results of the proposed OA-RVM model, focusing on prediction accuracy as evaluated by the root mean squared error (RMSE), presented in Table 1, and the correlation coefficient (COR) presented in Table 2.

Firstly, we observe that for both emotion dimensions and classes, OA-RVM outperforms RVM and SVM in terms of both COR and RMSE. The improvement is especially noticeable in terms of COR rather than RMSE. This can be justified by the fact that the goal of OA-RVM is to enforce common, temporal output patterns, thus increasing the covariance of the prediction with the ground truth. The prediction results provided by SVR and RVM are fairly similar, with RVM in general achieving better correlation with the ground truth. Given the above-mentioned results, we will focus our attention to the OA-RVM prediction results and compare them to the results we have previously presented in [21].

Focusing on the RMSE results of each class in isolation, we denote that for the positive class arousal appears to be easier to predict than valence. This is in agreement with the results presented in [21]. Nevertheless, for the same class, the COR achieved is higher for valence, showing that the structure of the valence dimension is modeled more accurately. When analyzing the results obtained for the negative class we observe that valence prediction is better than arousal prediction. In fact, considering the RMSE metric, arousal prediction for the

[4] Note that each sequence usually contains some portion of both positively/negatively *valenced* frames.

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005

Table 2
Evaluating SVM, RVM and OA-RVM in terms of COR for predicting arousal and valence from face, shoulder and audio cues. Results are averaged across four-fold subject-independent cross-validation. Best results are indicated in bold.

		Valence			Arousal		
Class	Cue	SVM	RVM	OA-RVM	SVM	RVM	OA-RVM
POS	Face	0.28	0.30	**0.43**	0.09	0.09	**0.16**
	Shoulders	0.01	0.16	**0.32**	0.12	0.19	**0.30**
	Audio	0.02	0.03	**0.19**	0.04	0.07	**0.21**
NEG	Face	0.14	0.20	**0.27**	0.13	0.18	**0.27**
	Shoulders	0.14	0.28	**0.29**	0.09	0.09	**0.22**
	Audio	0.01	0.05	**0.10**	0.23	0.23	**0.38**

negative class appears to be the most challenging case for the OA-RVM prediction framework.

Let us now compare the results obtained by employing different sets of nonverbal cues. When utilizing the facial expression cues, the correlation between the prediction and the ground truth appears to be equivalent for both emotion dimensions. In general, correlation obtained for the negative class appears to be highly dependent on the set of cues employed.

Related work on dimensional emotion recognition reported that arousal can be much better predicted than valence using audio cues [24,25,47]. Results obtained from our experiments are in line with such findings, showing that audio cues appear to provide the best prediction results (in terms of RMSE) for the arousal dimension. When considering the COR metric and the negative class, audio cues *again* appear to provide the best prediction results (0.38) compared to facial expression (0.27) and shoulder cues (0.22). For the positive class, while the audio cues still provide better correlation compared to using the facial expression cues, the shoulder cues appear very capable of capturing the arousal structure (and perform better than the audio cues).

It is well known that the facial expression cues are very informative for predicting valence. Our RMSE-based results confirm this, utilizing the facial expression cues provides better prediction results for the valence dimension. The shoulder cues also appear to be better at capturing useful information regarding the valence dimension compared to the audio cues.

When evaluating the *valence prediction models* in terms of the correlation metric, the models trained using the visual cues in general appear to perform better than the models trained using the audio cues (see Table 2). Additionally, for the negative class, the prediction models trained on the shoulder cues appear to slightly outperform the models trained on the facial expression cues.

In Fig. 5, we illustrate the average results for both classes evaluated in terms of RMSE and COR. Overall, we observe that regardless of the set of cues utilized or dimensions predicted, there is a significant increase in terms of correlation when applying OA-RVM. As denoted earlier, compared to OA-RVM and RVM, SVM provides the lowest correlation. Additionally, it can be seen that prediction models trained with facial expressions provide the lowest RMSE for the valence dimension, and the prediction models trained using the audio cues provide the lowest RMSE for the arousal dimension.

In Fig. 6 we also provide an illustrative comparison between the predictions generated by OA-RVM and RVM, on test data, with respect to the ground truth (utilizing different cues).

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 utilizing input and output associations as well as temporal context.

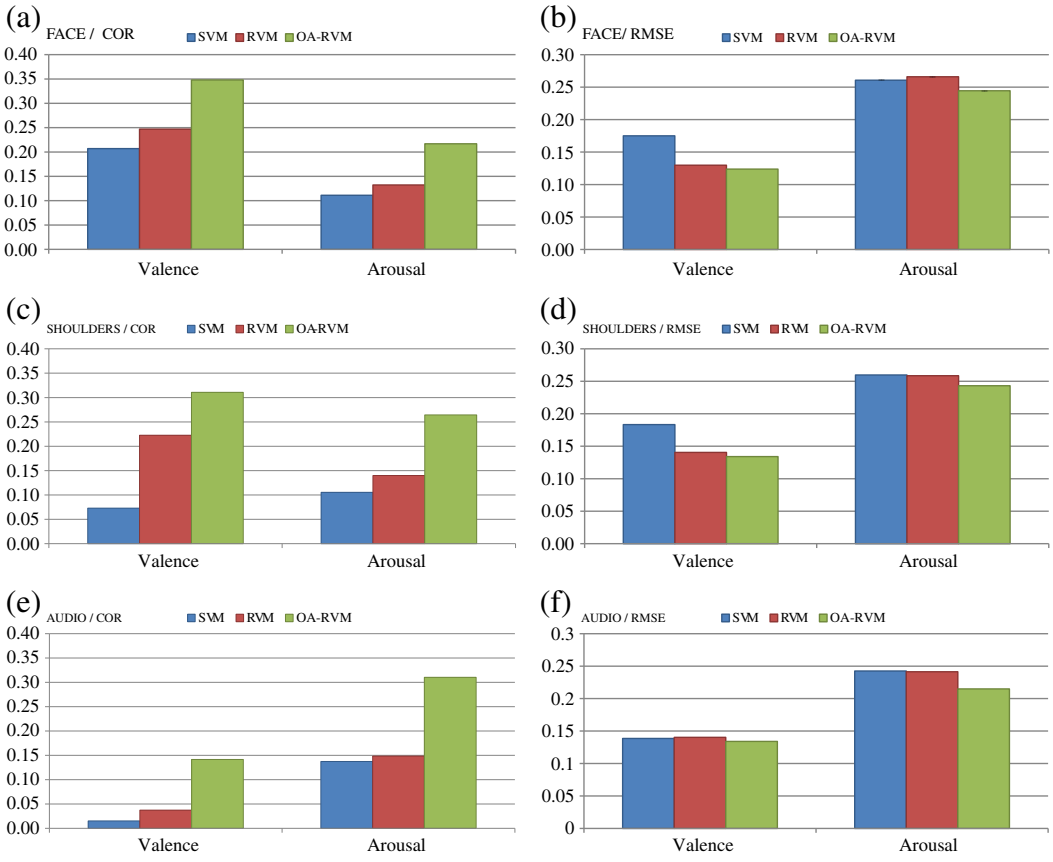

Fig. 5. Comparing the prediction COR (first column) and RMSE (second column) when utilizing (a,b) facial expressions, (c,d) shoulder movement and (e,f) audio cues.

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005

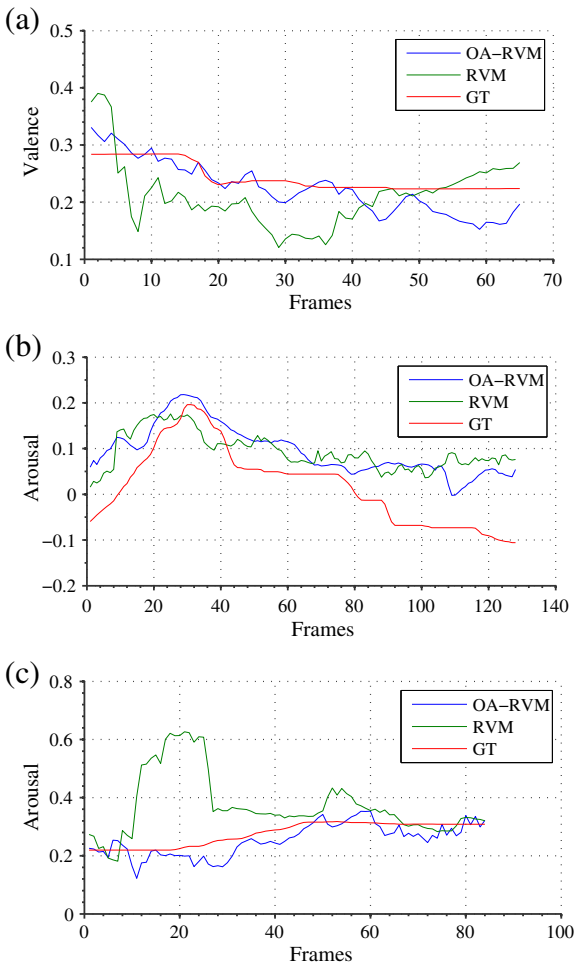

Fig. 6. An illustrative comparison between the predictions generated by OA-RVM and RVM, on test data, with respect to the ground truth (GT), utilizing different cues: (a) facial expressions, (b) shoulder movements, and (c) audio cues.

Psychological research findings suggest that there exist gender-related differences in expressing emotions (e.g., women appear to be more facially expressive than men [48]). In our experiments we found no consistent differentiations between male and female subjects. However, the data set employed is relatively small in order to draw generalizing conclusions regarding gender-related differences.

8. 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 the 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 heterogeneous nonverbal cues (facial expressions, shoulder movement and audio cues) and demonstrated its advantages and efficiency over a comprehensive set of experiments using subject-independent cross-validation. Our experimental results show that:

- OA-RVM outperforms both RVM and SVR in terms of prediction accuracy (RMSE) and prediction structure (COR), regardless of the set of cues utilized or emotion dimensions predicted. 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.
- 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 utilizing 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 generalization capability of the OA-RVM model over different data set(s) and subjects. It will also be interesting to investigate how (any) additional properties could be added to the proposed framework to fuse features coming from multiple heterogenous cues and modalities.

Acknowledgments

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

References

[1] H. Gunes, M. Pantic, Automatic, dimensional and continuous emotion recognition, Int. J. Synth. Emotions 1 (1) (2010) 68–99.
[2] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.
[3] J.A. Russell, A circumplex model of affect, J. Personal. Soc. Psychol. 39 (1980) 1161–1178.
[4] R. Lane, L. Nadel, Cognitive Neuroscience of Emotion, Oxford Univ. Press, 2000.
[5] A. Mehrabian, J. Russell, An Approach to Environmental Psychology, Cambridge, New York, 1974.
[6] A.M. Oliveira, M.P. Teixeira, I.B. Fonseca, M. Oliveira, Joint model-parameter validation of self-estimates of valence and arousal: probing a differential-weighting model of affective intensity, 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, Motiv. Emot. 21 (1997) 323–348.
[8] P.A. Lewis, et al., Neural correlates of processing valence and arousal in affective words, Cereb. Cortex 17 (3) (2007) 742–748 http://dx.doi.org/10.1093/cercor/bhk024, doi:10.1093/cercor/bhk024.
[9] J.P. Weston, O. Chapelle, A. Elisseeff, B. Scholkopf, V. Vapnik, Kernel Dependency Estimation, Tech. Rep. 98, Germany, August 2002.
[10] C. Cortes, M. Mohri, J. Weston, A general regression technique for learning transductions, Proc. of Int. Conf. on Machine Learning, ACM, New York, NY, USA, 2005, pp. 153–160.
[11] L. Bo, C. Sminchisescu, Structured output-associative regression, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 2009, pp. 2403–2410.
[12] L. Bo, C. Sminchisescu, Twin Gaussian processes for structured prediction, Int. J. Comput. Vision 87 (2010) 28–52.
[13] K. Mitra, A. Veeraraghavan, R. Chellappa, Robust RVM regression using sparse outlier model, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 2010, pp. 1887–1894.
[14] A. Thayananthan, R. Navaratnam, B. Stenger, P.H.S. Torr, R. Cipolla, Multivariate relevance vector machines for tracking, Proc. of European Conference on Computer Vision, 2006, pp. 124–138.
[15] H. Gunes, B. Schuller, M. Pantic, R. Cowie, Emotion representation, analysis and synthesis in continuous space: a survey, Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 827–834.
[16] M.A. Nicolaou, H. Gunes, M. Pantic, Audio-visual classification and fusion of spontaneous affective data in likelihood space, Proc. of Int. Conf. on Pattern Recognition, 2010, pp. 3695–3699.
[17] D. McDuff, R. El Kaliouby, K. Kassam, R. Picard, Affect valence inference from facial action unit spectrograms, Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2010, pp. 17–24.
[18] D. Glowinski, A. Camurri, G. Volpe, N. Dael, K. Scherer, Technique for automatic emotion recognition by body gesture analysis, Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–6.
[19] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, S. Kollias, Emotion recognition through facial expression analysis based on a neurofuzzy method, J. Neural Networks 18 (2005) 423–435.
[20] H. Gunes, M. Pantic, Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners, Proc. Int. Conf. on Intelligent Virtual Agents, 2010, pp. 371–377.

[21] M.A. Nicolaou, H. Gunes, M. Pantic, Output-associative RVM regression for dimensional and continuous emotion prediction, Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 16–23.
[22] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic emotion recognition: a benchmark comparison of performances, Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, 2009, pp. 552–557.
[23] M. Wöllmer, B. Schuller, F. Eyben, G. Rigoll, Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening, IEEE J. Sel. Top. Sign. Proces. 4 (5) (2010) 867–881.
[24] M. Wollmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, Abandoning emotion classes — towards continuous emotion recognition with modelling of long-range dependencies, Proc. of INTERSPEECH Conference, 2008, pp. 597–600.
[25] M. Grimm, K. Kroschel, Emotion estimation in speech using a 3d emotion space concept, Proc. IEEE Automatic Speech Recognition and Understanding Workshop, 2005, pp. 381–385.
[26] M. Grimm, K. Kroschel, S. Narayanan, The Vera am Mittag German audio-visual emotional speech database, Proc. of IEEE Int. Conf. on Multimedia & Expo, 2008, pp. 865–868.
[27] H. Espinosa, C. Garcia, L. Pineda, Features selection for primitives estimation on emotional speech, Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing, 2010, pp. 5138–5141.
[28] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, K. Karpouzis, Modeling naturalistic affective states via facial and vocal expressions recognition, Proc. of ACM Int. Conf. on Multimodal Interaction, 2006, pp. 146–154.
[29] I. Kanluan, M. Grimm, K. Kroschel, Audio-visual emotion recognition using an emotion recognition space concept, Proc. of European Signal Processing Conference, 2008.
[30] M. Wöllmer, A. Metallinou, F.E. Schuller, S. Narayanan, Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling, Proc. of INTERSPEECH Conference, 2010, pp. 2362–2365.
[31] F. Eyben, M. Wollmer, M.F. Valstar, H. Gunes, B. Schuller, M. Pantic, String-based audiovisual fusion of behavioural events for the assessment of dimensional affect, Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 322–329.
[32] A. Metallinou, A. Katsamanis, Y. Wang, S. Narayanan, Tracking changes in continuous emotion states using body language and prosodic cues, Proc. of IEEE Int. Conf. on Acoustics Speech and Signal Processing, 2011, pp. 2288–2291.
[33] S. Gilroy, M. Cavazza, M. Niiranen, E. Andre, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, M. Billinghurst, Pad-based multimodal affective fusion, Proc. of Int. Conf. on Affective Computing and Intelligent Interaction Workshops, 2009, pp. 1–8.
[34] M.A. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space, IEEE Trans. Affect. Comput. 2 (2) (2011) 92–105.
[35] H. Gunes, M.A. Nicolaou, M. Pantic, Computer analysis of human behavior, Ch. Continuous Analysis of Affect from Voice and Face, Springer-Verlag, 2011, pp. 255–291.
[36] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (2001) 211–244.
[37] M.E. Tipping, A. Faul, Fast marginal likelihood maximisation for sparse Bayesian models, Proc. of Int. Workshop on Artificial Intelligence and Statistics, 2003, pp. 3–6.
[38] D.H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.
[39] L. Breiman, Stacked regressions, Mach. Learn. 24 (1996) 49–64http://dx.doi.org/10.1007/BF00117832, doi:10.1007/BF00117832 URL.
[40] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, K. Karpouzis, The humane database: addressing the needs of the affective computing community, Proc. of Int. Conf. on Affective Computing and Intelligent Interaction, 2007, pp. 488–500.
[41] M.A. Nicolaou, H. Gunes, M. Pantic, Automatic segmentation of spontaneous data using dimensional labels from multiple coders, Proc. of LREC Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, 2010, pp. 43–48.
[42] I. Patras, M. Pantic, Particle filtering with factorized likelihoods for tracking facial features, Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2004, pp. 97–102.
[43] M.K. Pitt, N. Shephard, Filtering via simulation: auxiliary particle filters, J. Am. Stat. Assoc. 94 (446) (1999) 590–616.
[44] D. Jurafsky, J.H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd Edition Prentice Hall, 2008.
[45] B. Paul, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, Proc. of the Institute of Phonetic Sciences, 1993, pp. 97–110.
[46] S. Petridis, H. Gunes, S. Kaltwang, M. Pantic, Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities, Proc. of ACM Int. Conf. on Multimodal Interfaces, 2009, pp. 23–30.
[47] K.P. Truong, D.A. Leeuwen van, M.A. Neerincx, F.M. Jong de, Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion, Proc. of INTERSPEECH Conference, 2009, pp. 2027–2030.
[48] A.M. Kring, A.H. Gordon, Sex differences in emotion: expression, experience, and physiology, J. Personal. Soc. Psychol. 74 (3) (1998) 686–703.

Please cite this article as: M.A. Nicolaou, et al., Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. (2012), doi:10.1016/j.imavis.2011.12.005