World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:3, 2016

Multivariate Output-Associative RVM for Multi-Dimensional Affect Predictions

Achut Manandhar, Kenneth D. Morton, Peter A. Torrione, Leslie M. Collins

Abstract—The current trends in affect recognition research are to consider continuous observations from spontaneous natural interactions in people using multiple feature modalities, and to represent affect in terms of continuous dimensions, incorporate spatio-temporal correlation among affect dimensions, and provide fast affect predictions. These research efforts have been propelled by a growing effort to develop affect recognition system that can be implemented to enable seamless real-time human-computer interaction in a wide variety of applications. Motivated by these desired attributes of an affect recognition system, in this work a multi-dimensional affect prediction approach is proposed by integrating multivariate Relevance Vector Machine (MVRVM) with a recently developed Output-associative Relevance Vector Machine (OARVM) approach. The resulting approach can provide fast continuous affect predictions by jointly modeling the multiple affect dimensions and their correlations. Experiments on the RECOLA database show that the proposed approach performs competitively with the OARVM while providing faster predictions during testing.

Keywords—Dimensional affect prediction, Output-associative RVM, Multivariate regression.

I. Introduction

ANALYZING affective human behavior is an important aspect for developing affect sensitive systems that have a wide variety of applications in human-computer interaction [1], [2], clinical and biomedical studies [3], [4], autism-related assistive technology [5], adaptive learning environments [6], affect recognition in the car [7], multimedia [8], [9], and entertainment [10]. All of these applications seek real-time continuous human-human or human-computer interaction. A driver assistive system needs to react immediately to a drowsy driver [7] and an autism-related assistive technology needs to provide real-time affect recognition to enable effective communication and avoid undesired intervention [5]. Thus the current trend in affect recognition is to develop a fast system that can provide real-time feedback, enabling seamless interaction [1], [11], [12]. The need to operate in real-time also drives the need to process continuous input signals to analyze affect continuously, which has motivated several recent affect recognition approaches that consider temporal data [7], [13], [14]. In addition to the continuity in the input signals, many recent efforts have represented affect itself in continuous dimensional space to overcome the limitations of category-specific representation in modeling complex human emotion [15], [16], [17], [18]. Moreover, recent efforts have shown these continuous affect dimensions to be correlated with each other [15], [16], [19].

Achut Manandhar, Kenneth D. Morton, Peter A. Torrione, and Leslie M. Collins are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708 USA (e-mail: lcollins@duke.edu).

Although the current trend has been to develop approaches that can model continuous input observations, a vast majority of approaches have focused on using techniques that assume observations are independent. Many of these efforts use Support Vector Machines (SVMs) [4], Support Vector Regression (SVR) [2], [20], and Relevance Vector Machines (RVMs) [8], none of which incorporate temporal correlation. In order to incorporate past and future observations, Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) [7], [21] and Bidirectional Long-Short Term RNNs (BLSTM-RNNs) [22], [21] have been implemented for affect prediction. These approaches overcome the static RNN's limitations by allowing the network to store and retreive information over long periods of time. They can learn the amount of context instead of manually defining fixed-size temporal windows [22]. Moreover, they can be implemented to model affect dimensions simultaneously [21]. The LSTM-RNN approaches have been shown to outperform the static RNN and the standard SVR approaches for affect recognition [22], [21], [23]. Along the same vein, in order to model correlation between affect dimensions, a multi-layered hybrid framework [13] has been developed, where in the first layer, a LSTM is used to generate continuous arousal and valence estimates. These estimates are used in the second layer by an Auto-Regressive Coupled Hidden Markov Model (ACHMM) to capture the correlation between the affect dimensions, both of which are used in the third layer for the final classfiication using a SVM. Inspired by this framework, a two-stage approach has been used in the OARVM [14] to model the correlation between arousal and valence dimensions, where in the first stage, two independent RVMs are used to obtain continuous affect estimates. These estimates along with the original input observations are used by two new RVMs in the second stage to generate the final arousal and valence estimates. Using this two-stage approach, the OARVM can model both the temporal dependencies as well as dependencies between the affect dimensions. By incorporating correlation, the OARVM has been shown to outperform the traditional RVM and SVR approaches for affect prediction [14].

Despite fulfilling several requirements of an affect recognition approach, the OARVM requires training independent regressors for each affect dimension in each stage of the learning process. Predicting each affect dimension separately may translate to additional computation time during testing. The testing time of the OARVM increases approximately linearly with the number of affect dimensions to be predicted. Although the OARVM is specifically applied to predict the most widely used [13], [22], [21] arousal and valence dimensions, it can be applied to predict multiple

affect dimensions. Many efforts have considered additional affect dimensions to model human emotion [16], [17], [18] and there is ongoing research in determining a useful number of affect dimensions [17], [18], [12]. A possible solution to generalize the OARVM to predict multiple affect dimensions without increasing the testing time may be to implement multiple regressors in parallel in each stage. An alternative solution, proposed in this work, is to jointly model multiple affect dimenions by using a multivariate RVM (MVRVM) [24]. The resulting approach models correlation among multiple affect dimensions simultaneously, enabling fast affect predictions during testing. Our experiments on the RECOLA database [25], [21] show that the proposed approach performs competitively with the OARVM while reducing the prediction time during testing.

The remainder of this paper is organized as follows. Section II describes different types of RVM-based approaches that can be used for dimensional affect prediction problem, including the proposed approach. Section III evaluates the RVM-based approaches on the RECOLA database. Finally, Section IV concludes this paper and discusses avenues for future work.

II. RVM Models for Dimensional Affect Prediction

As mentioned in Section I, considering each affect dimension independently and assuming temporal independence, a collection of continuous affect dimensions and their corresponding observations can be used to learn a separate regression function for each affect dimension. However, the affect dimensions are known to be both correlated in time and among themselves [15], [16], [19]. With this motivation, the OARVM [14] extends the RVM by modeling the input-output association between affect dimensions. However, the OARVM learns a separate regression function per affect dimension, increasing testing time as more affect dimensions are predicted. The proposed approach further extends the OARVM by simultaneously learning multiple affect dimensions, enabling fast testing time while also modeling their input-output associations.

A. RVM

The RVM [26] is a Bayesian sparse kernel technique for regression and classification that shares many characteristics with the SVM [27] while avoiding its limitations. In contrast to the SVM, the RVM typically learns a much sparser model while maintaining a comparable accuracy, and unlike the SVM, the RVM can provide probabilistic predictions.

Similar to the SVM, given a collection of input/ouput pairs, $\{\boldsymbol{x}_n, t_n\}, n \in \{1, ..., N\}$, where $\boldsymbol{x}_n$ is the n^{th} input observation and t_n is the corresponding output response, in the RVM, the output responses are assumed to be generated from a linear model with added Gaussian noise as

$$\begin{aligned} \boldsymbol{t}|\boldsymbol{w}, \beta &\sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{w}, \beta^{-1}\boldsymbol{I}_N) && (1)\\ \boldsymbol{t} &= [t_1, ..., t_N]^T && (2)\\ \boldsymbol{\Phi} &= [\boldsymbol{\phi}(\boldsymbol{x}_1), ..., \boldsymbol{\phi}(\boldsymbol{x}_N)]^T && (3)\\ \boldsymbol{\phi}(\boldsymbol{x}_n) &= [1, \phi_1(\boldsymbol{x}_n), ..., \phi_M(\boldsymbol{x}_n)]^T && (4)\\ &= [1, k(\boldsymbol{x}_n, \boldsymbol{x}_1), ..., k(\boldsymbol{x}_n, \boldsymbol{x}_M)]^T, \end{aligned}$$

where β is the precision on the noise, $\boldsymbol{\Phi}$ is the $[N \times \text{N+1}]$ design matrix, $\boldsymbol{\phi}(\boldsymbol{x}_n)$ is the vector of basis functions defined over the input observations, and $\boldsymbol{w}^T = [w_0, w_1, ..., w_M]^T$ is the corresponding vector of weights. In order to promote sparsity, the weights are assumed to be drawn from zero-mean Gaussians as

$$\boldsymbol{w}|\boldsymbol{\alpha} = \prod_{m=0}^{M} \mathcal{N}(w_m|0, \alpha_m^{-1}), \quad (5)$$

where α_m is the precision on the weight w_m. The data generation process is detailed in Fig. 1(a) and forms the basis for the extensions of the RVM described in this work.

The posterior density on the weights can be analytically computed as [26]:

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{\alpha}, \beta) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) && (6)\\ \boldsymbol{\Sigma} &= (\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{A})^{-1} && (7)\\ \boldsymbol{\mu} &= \beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{t}, && (8) \end{aligned}$$

where $\boldsymbol{A} = \text{diag}(\alpha_0, ..., \alpha_M)$. To estimate β and other hyperparameters, various approximation approaches have been formulated [28], [26], [29]. In this work, due to the need to perform fast inference and the availability of large amount of training data, the fast sequential parameter estimation approach [29] has been adopted [1].

Having estimated the model parameters, given a new observation, $\boldsymbol{x}_*$, the posterior predictive density of the output response, t_*, can be approximated at the maximum likelihood estimates, $\{\boldsymbol{\alpha}_{ML}, \beta_{ML}\}$ as [29]:

$$\begin{aligned} p(t_*|\boldsymbol{\alpha}_{ML}, \beta_{ML}) &= \mathcal{N}(t_*|y_*, \sigma_*^2), && (9)\\ y_* &= \boldsymbol{\mu}^T\boldsymbol{\phi}(\boldsymbol{x}_*) && (10)\\ \sigma_*^2 &= \beta_{ML}^{-1} + \boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x}_*), && (11) \end{aligned}$$

where the posterior predictive mean, y_*, is the basis vector weighted by a sparse vector of mean weights, $\boldsymbol{\mu}$, thus resulting in a sparse representation of data; and σ_*^2 is the variance on the predictions.

As described earlier, by representating the independently generated output response as an affect dimension, a RVM can be used to learn a separate regression function for each affect dimension. The RVM is extended in the next section to incorporate both temporal correlation as well as correlation among multiple affect dimensions.

[1]Tipping's SparseBayes MATLAB®software was used for implementing the RVM, which is publicly available at http://www.miketipping.com/downloads.htm

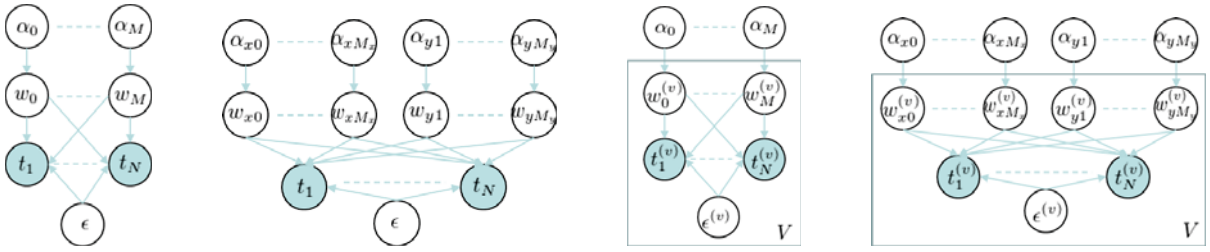

Fig. 1 Graphical models of (a) RVM [26], (b) OARVM [14], (c) MVRVM [24], and (d) the proposed MVOARVM: The OARVM extends the RVM by using the output estimates as additional input observations, which is reflected in the graphical model as the additional weight vector $[w_{y1}, .., w_{yM_y}]^T$ and the corresponding precision vector $[\alpha_{y1}, .., \alpha_{yM_y}]^T$. The MVRVM extends the RVM by simultaneously modeling multiple output variables, which is reflected by the addition of a plate that replicates over the total number of variables V. Finally, the proposed MVOARVM combines the former two by incorporating both output-association as well as multivariate regression

B. Output-Associative RVM

To overcome the limitations of the RVM in modeling continuous correlated affect dimensions, a new approach called the OARVM [14] has been recently developed to model the inherent spatio-temporal dependencies of arousal and valence dimensions. The OARVM extends the standard RVM by learning the non-linear input-output dependencies in a two-stage process. In the first stage, separate output estimates of each affect dimension are obtained using independent regression functions. These output estimates can be obtained using a RVM or any other standard regression technique. In the second stage, the output estimates of both arousal and valence, spanning a pre-defined temporal window, along with the original input observations are then collectively used to learn a new RVM for each affect dimension.

Extending the standard RVM to the OARVM only requires minor modification in the model described in Section II-A. Since the OARVM makes predictions based on both input observations and output estimates, the linear model, $\boldsymbol{\Phi w}$, in 1 is expressed as a linear combination of two different design matrices as

$$\boldsymbol{t}|\boldsymbol{w}, \beta \quad \sim \quad \mathcal{N}(\boldsymbol{\Phi w}, \beta^{-1}\boldsymbol{I}_N) \tag{12}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_x|\boldsymbol{\Phi}_y]$ is the $[N \times (M_x + M_y)]$ concatenated design matrix and $\boldsymbol{w} = [\boldsymbol{w}_x|\boldsymbol{w}_y]^T$ is the corresponding concatenated weight vectors. The augmented design matrix $\boldsymbol{\Phi}$ is defined as

$$\begin{aligned}
\boldsymbol{\Phi}_x &= [\boldsymbol{\phi}(\boldsymbol{x}_1), ..., \boldsymbol{\phi}(\boldsymbol{x}_N)]^T && (13)\\
\boldsymbol{\phi}(\boldsymbol{x}_n) &= [1, k_x(\boldsymbol{x}_n, \boldsymbol{x}_1), ..., k_x(\boldsymbol{x}_n, \boldsymbol{x}_{M_x})]^T && (14)\\
\boldsymbol{\Phi}_y &= [\boldsymbol{\phi}_y(\boldsymbol{y}_1), ..., \boldsymbol{\phi}_y(\boldsymbol{y}_N)]^T && (15)\\
\boldsymbol{\phi}_y(\boldsymbol{y}_n) &= [k_y(\boldsymbol{y}_n, \boldsymbol{y}_1), ..., k_y(\boldsymbol{y}_n, \boldsymbol{y}_{M_y})]^T, && (16)
\end{aligned}$$

where $\boldsymbol{y}_n$ is a vector of multidimensional output estimates defined over a temporal window, $\boldsymbol{\phi}(\boldsymbol{x}_n)$ and $\boldsymbol{\phi}(\boldsymbol{y}_n)$ are the vectors of basis functions defined over the input observations and the output estimates respectively, and finally, $\boldsymbol{w}_x = [w_{x0}, w_{x1}, ..., w_{xM_x}]^T$ and $\boldsymbol{w}_y = [w_{y1}, ..., w_{yM_y}]^T$ are their corresponding weight vectors.

The sparsity promoting priors on the weight vectors are similar to 5 and are defined as

$$\begin{aligned}
\boldsymbol{w}_x|\boldsymbol{\alpha}_x &= \prod_{m=0}^{M_x} \mathcal{N}(w_{xm}|0, \alpha_{xm}^{-1}) && (17)\\
\boldsymbol{w}_y|\boldsymbol{\alpha}_y &= \prod_{m=0}^{M_y} \mathcal{N}(w_{ym}|0, \alpha_{ym}^{-1}), && (18)
\end{aligned}$$

where α_{xm} and α_{ym} are the precision on the weights w_{xm} and w_{ym} respectively. The generative process of the OARVM is detailed in Fig. 1(b), which differs from Fig. 1(a) only in terms of the additional weight vectors and their corresponding precision vectors. Not surprisingly, the posterior density of the weights and the posterior predictive density of a new observation are similar to 7, 8, 10, and 11 with their notations redefined as

$$\begin{aligned}
\boldsymbol{\Sigma} &= [\boldsymbol{\Sigma}_{xx}, \boldsymbol{\Sigma}_{xy}; \boldsymbol{\Sigma}_{xy}, \boldsymbol{\Sigma}_{yy}] && (19)\\
\boldsymbol{\mu} &= [\boldsymbol{\mu}_x; \boldsymbol{\mu}_y] && (20)\\
\boldsymbol{A} &= \text{diag}(\alpha_{x0}, ..., \alpha_{xM_x}, \alpha_{y1}, ..., \alpha_{yM_y}), && (21)
\end{aligned}$$

where $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yy}$, and $\boldsymbol{\Sigma}_{xy}$ are the covariances of the weight vectors $\boldsymbol{w}_x$, $\boldsymbol{w}_y$, and between $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ respectively. Likewise, $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are the means of the weight vectors $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ respectively. Thus by exploiting the input-output association between valence and arousal dimensions, the OARVM learns a separate regression function for each affect dimension. Similar to several other approaches [13], [22], [21], the OARVM also considers only the two most commonly used affect dimensions - arousal and valence. However, other efforts have considered one or more of the additional affect dimensions [16], [17], [18] and there is an ongoing research towards determining a useful number of affect dimensions to model human emotion [17], [18], [12].

In theory, the OARVM can model multiple affect dimensions but it learns a separate regression function for each affect dimension, and thus requires keeping track of separate relevance vectors and other model parameters per regressor. Consequently, the testing time increases approximately linearly with the number of affect dimensions to be predicted. As described in Section I, one of the crucial attributes of an affect recognition system is to be able to

provide immediate feedback [1], [11], [12], which among other factors also depends on the testing time of the affect recognition approach. In order to speed-up the OARVM's testing time, one possible solution may be to run multiple independent RVMs in parallel during each stage of the learning process. Another alternative, proposed in this work, is to simultaeously predict multiple affect dimensions, keeping the testing time constant irrespective of the number of affect dimensions.

The following section first describes the multivariate RVM, which is then combined with the OARVM to develop a new approach for predicting multiple affect dimensions.

C. Multivariate RVM

In order to model muliple continuous output variables, a multivarite Relevance Vector Machine (MVRVM) [24] has been developed to estimate the 3D pose of an object from a single-view camera continuously on a frame-by-frame basis during tracking. In this extension of the RVM, common input observations are used to learn common sparse relevance vectors to predict multiple output variables simultaneously.

Consider a collection of observations and multiple output responses, $\{\boldsymbol{x}_n, \boldsymbol{t}_n\}, n \in \{1, ..., N\}$, where the n^{th} output response, $\boldsymbol{t}_n = [t_n^{(1)}, ..., t_n^{(V)}]$, represents V different output variables. Assuming the output variables are independent, the generative process for MVRVM can be described as

$$\begin{aligned} \boldsymbol{t}|\boldsymbol{W}, \boldsymbol{\beta} &= \prod_{v=1}^{V} \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{w}^{(v)}, \boldsymbol{I}_N/\beta^{(v)}); v \in \{1, ..., V\} && (22)\\ \boldsymbol{t} &= [\boldsymbol{t}^{(1)}, ..., \boldsymbol{t}^{(V)}] && (23)\\ \boldsymbol{\Phi} &= [\boldsymbol{\phi}(\boldsymbol{x}_1), ..., \boldsymbol{\phi}(\boldsymbol{x}_N)]^T && (24)\\ \boldsymbol{W} &= [\boldsymbol{w}^{(1)}, ..., \boldsymbol{w}^{(V)}], && (25) \end{aligned}$$

where $\boldsymbol{\Phi}$ is the common $[N \times M]$ design matrix for all output variables, $\boldsymbol{W}$ is the $[M \times V]$ weight matrix, $\boldsymbol{t}^{(v)} = [t_1^{(v)}, ..., t_N^{(v)}]^T$, $\boldsymbol{w}^{(v)} = [w_0^{(v)}, ..., w_M^{(v)}]^T$, and $\beta^{(v)}$ are respectively the output response, the weight vector, and the noise precision corresponding to the v^{th} output variable. Assuming the weight vectors across variables are independent, the sparsity promoting priors on the weight vectors can be expressed as

$$\boldsymbol{W}|\boldsymbol{\alpha} = \prod_{v=1}^{V} \prod_{m=0}^{M} \mathcal{N}(w_m^{(v)}|0, \alpha_m^{-1}) = \prod_{v=1}^{V} \mathcal{N}(\boldsymbol{w}^{(v)}|0, \boldsymbol{A}^{-1}), \quad (26)$$

where the weight vectors for all variables share the common precision $\boldsymbol{A}$, thus resulting in common relevance vectors. The data generation process in the MVRVM is detailed in Fig. 1(c), which differs from Fig. 1(a) by simply allowing the output variable, the weight vector, and the precision on noise to replicate over V variables.

Once again, the posterior density of the weights are similar to 7 and 8, with only slight difference in notations to account for multiple variables as

$$\begin{aligned} \boldsymbol{\Sigma}^{(v)} &= (\beta^{(v)}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{A})^{-1} && (27)\\ \boldsymbol{\mu}^{(v)} &= \beta^{(v)}\boldsymbol{\Sigma}^{(v)}\boldsymbol{\Phi}^T\boldsymbol{t}^{(v)}. && (28) \end{aligned}$$

The key benefit of the MVRVM over using multiple RVMs is that given a new observation, $\boldsymbol{x}_*$, the posterior predictive density for the output variables, $\boldsymbol{t}_*$, can be estimated simultaneously as follows

$$\begin{aligned} \boldsymbol{y}_* &= \boldsymbol{\mu}^T\boldsymbol{\phi}(\boldsymbol{x}_*) && (29)\\ (\sigma_*^{(v)})^2 &= 1/\beta_{ML}^{(v)} + \boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{\Sigma}^{(v)}\boldsymbol{\phi}(\boldsymbol{x}_*), && (30) \end{aligned}$$

where $\boldsymbol{y}_* = [y_*^{(1)}, ..., y_*^{(V)}]^T$ is a vector of multivariate output estimates, $\boldsymbol{\mu} = [\boldsymbol{\mu}^{(1)}, ..., \boldsymbol{\mu}^{(V)}]$ is a $[M \times V]$ matrix of weights, and $(\sigma_*^{(v)})^2$ is the variance on the estimate of the v^{th} variable.

By representing the multivariate output response, $\boldsymbol{t}_n$, as multiple affect dimensions at the n^{th} instance in time, the MVRVM can be used to model multiple affect dimensions, however the MVRVM does not model correlation among affect dimensions. Motivated by the continuous affect prediction problem and capitalizing on the benefits of both OARVM and MVRVM, a new approach is proposed in the next section to model continuous correlated multiple variables.

D. Multivariate Output-Associate RVM

The OARVM extends the RVM to incorporate correlation in continuous multiple output variables but learns separate regression function for each output variable, incurring additional computational cost during testing. In contrast, the MVRVM extends the RVM to model multiple variables simultaneously but fails to incorporate correlation among output variables. Since an affect recognition system should ideally model the correlation among affect dimensions while providing fast predictions, in this work both the OARVM and the MVRVM are combined to develop a new approach called Multivariate OARVM (MVOARVM) that simultaneously models multiple variables while also incorporating their correlations.

Not surprisingly, the proposed model is very similar to the ones detailed in Sections II-B and II-C with only small modifications. The data generation process is similar to 22 and is described as follows

$$\boldsymbol{t}|\boldsymbol{W}, \boldsymbol{\beta} = \prod_{v=1}^{V} \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{w}^{(v)}, \boldsymbol{I}_N/\beta^{(v)}); v \in \{1, ..., V\} \quad (31)$$

where the multivariate output matrix, $\boldsymbol{t}$, and the weight matrix, $\boldsymbol{W}$, are defined similarly as in 23 and 25, the design matrix, $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_x|\boldsymbol{\Phi}_y]$, is defined similarly as in 13-16, and $\beta^{(v)}$ is the noise precision corresponding to the v^{th} variable. Moreover, the design matrix is common to all output variables.

In contrast to Sections II-B and II-C, each of the weight vectors, $\boldsymbol{w}^{(v)} = [\boldsymbol{w}_x^{(v)}|\boldsymbol{w}_y^{(v)}]^T$, is defined as the variable-specific concatenated vectors, where $\boldsymbol{w}_x^{(v)} = [w_{x0}^{(v)}, w_{x1}^{(v)}, ..., w_{xM_x}^{(v)}]^T$ and $\boldsymbol{w}_y^{(v)} = [w_{y1}^{(v)}, ..., w_{yM_y}^{(v)}]^T$ are the weight vectors corresponding to the basis functions defined over the input observations and the output estimates respectively. With the weight vectors redefined, their prior density is similar to 26. The data generation process is further detailed in Fig. 1(d), which combines Figs. 1(b) and 1(c) in two steps - first, by adding the additional weight and precision vectors corresponding to the basis functions defined

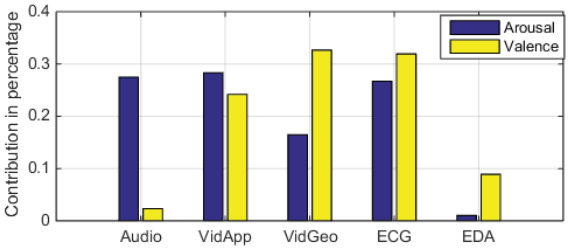

Fig. 2 Fraction of contribution of each feature modality used to predict the affect dimensions. The values are obtained by normalizing the linear regression weights for each affect dimension obtained with decision-fusion

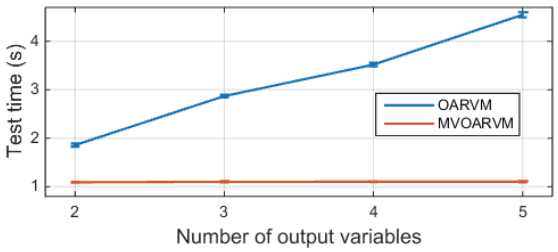

Fig. 3 Estimated test time as a function of the number of output variables to be predicted when using the video-geometric features. The values represent the mean over 10 different trials and the error bars represent 99% confidence interval. The results are similar for other features. The timings are based on a MATLAB®implementation using Intel®Core(TM) i7-2600 CPU 3.4 GHz

over the output estimates, and second, by replicating the output variables and weight vectors over V variables.

The posterior density of the weight vectors is similar to 27 and 28 with the notations redefined as

$$\boldsymbol{\Sigma}^{(v)} = [\boldsymbol{\Sigma}_{xx}^{(v)}, \boldsymbol{\Sigma}_{xy}^{(v)}; \boldsymbol{\Sigma}_{xy}^{(v)}, \boldsymbol{\Sigma}_{yy}^{(v)}] \quad (32)$$

$$\boldsymbol{\mu}^{(v)} = [\boldsymbol{\mu}_{x}^{(v)}; \boldsymbol{\mu}_{y}^{(v)}], \quad (33)$$

where $\boldsymbol{\Sigma}_{xx}^{(v)}$, $\boldsymbol{\Sigma}_{yy}^{(v)}$, and $\boldsymbol{\Sigma}_{xy}^{(v)}$ are the covariances of the weight vectors $\boldsymbol{w}_{x}^{(v)}$, $\boldsymbol{w}_{y}^{(v)}$, and between $\boldsymbol{w}_{x}^{(v)}$ and $\boldsymbol{w}_{y}^{(v)}$ respectively. Likewise, $\boldsymbol{\mu}_{x}^{(v)}$ and $\boldsymbol{\mu}_{y}^{(v)}$ are the means of the weight vectors $\boldsymbol{w}_{x}^{(v)}$ and $\boldsymbol{w}_{y}^{(v)}$ respectively. The precision matrix, A, is common to all variables and is still given by 21. With the notations redefined, the posterior predictive density is similar to 29 and 30. In this way, by adopting a similar process used by the MVRVM to extend the RVM to jointly model multiple variables, the proposed approach extends the OARVM to jointly model continuous correlated multiple affect dimensions while providing fast predictions during testing. The proposed approach integrates several important attributes of an affect recognition system described in Section I.

The different RVM-based approaches described in Sections II-A-II-D are evaluated using the RECOLA database [25], [21] in the following section.

III. Experiments

Experimental evaluation was performed on the RECOLA database [25], [21] used in the AV+EC 2015 challenge [23]. Using the database, the proposed approach is evaluated against the previous RVM-based approaches as well as the best-performing SVR+NN approach implemented in the challenge baseline paper [23]. The SVR+NN approach is the decision fusion of a SVR and a Neutral Network (NN), where the parameters of the SVR and the parameters and the architectures (feed-forward, LSTM, and BLSTM) of the NN are optimized on the development partition [23]. The performance of the approaches are first compared using individual feature modalities independently, followed by their decision-fusion. Similar to [23], the decision fusion was done by training a linear regression model on the regression outputs obtained using individual feature modalities on the development set. Following [23], [21], the Concordance Correlation Coefficient (CCC) [31] is used as a performance metric, which combines the Pearson's Correlation Coefficient and the mean square error in a single metric as

$$\text{CCC}(X,Y) = \frac{2 \times \text{COV}(X,Y)}{\text{VAR}(X) + \text{VAR}(Y) + (\text{E}(X) - \text{E}(Y))^2}, \quad (34)$$

where, in the context of dimensional affect prediction, X and Y represent equal length gold standard ratings and predictions for a particular affect dimension.

A. RECOLA Database

The RECOLA database [25], [21] is a corpus of naturalistic interactions generated in the context of remote collaborative work. The database consists of audio, video, and physiological features and dimensional affect ratings in terms of arousal and valence by six different raters along with the gold standard ratings. For the purpose of evaluation, the database has been equally divided into three partitions - training, development, and test. Readers are referred to the prior reports [25], [21], [23] to learn more about the database.

B. Data Preprocessing and Parameter Settings

Following [23], all feature sets are individually normalized per subject using a *z-score*. For the RVM-based approaches, one out of every twenty frames from the training set was considered to reduce the computation time. All frames were considered from the development and the test sets.

For the RVM-based approaches, the kernel width was optimized for each regression function on the development set using values in ranges $[10 - 50]$ and $[0.1 - 0.5]$ for the kernel functions corresponding to the input observations and the output estimates respectively. Following the Bayesian specification, the kernel width can also be learned as a model parameter, which is included as part of our future work. For the OARVM-based approaches, following [14], for each feature modality, the temporal window size was optimized on the development set using values in the range $[1 - 4]$ temporal steps in the downsampled training set, which corresponds to $[0.8 - 3.2]s$. The range lies within the ones used in [21], where the window size was optimized using the values in the range $[0.48 - 6.24]s$ for individual feature modalities and affect dimensions. The results in [21] suggest longer window size is better for valence compared to arousal. There is actually ongoing research in determining the best length of temporal window for a given feature modality and affect dimension [18], [12], [21]. In this work, for each feature modality, we

TABLE I
PERFORMANCE COMPARISON OF THE APPROACHES ON THE DEVELOPMENT SET WITH AUDIO, VIDEO (APPEARANCE AND GEOMETRIC), ECG, AND EDA FEATURE MODALITIES

	AROUSAL					VALENCE				
MODALITY	SVR+NN	RVM	mvRVM	oaRVM	mvoaRVM	SVR+NN	RVM	mvRVM	oaRVM	mvoaRVM
D-Audio	.287	.131	.120	**.474**	.333	.069	.060	.047	.081	**.108**
D-VApp	.103	.125	.112	.287	**.307**	.273	.252	.221	.446	.435
D-VGeo	.231	.124	.094	.229	.211	.325	.291	.263	**.510**	.476
D-ECG	.275	.185	.187	**.293**	.201	.183	.176	.173	.274	.272
D-EDA	.078	.055	.051	.085	**.109**	.204	.157	.147	**.232**	.216

The results for the decision-fusion of SVR+NN are obtained from the baseline paper [23]. For each affect dimension and feature modality, the numbers in bold indicate significantly better than the other approaches at $p < .01$ based on the Fisher Z-transform test [30].

TABLE II
PERFORMANCE COMPARISON OF THE APPROACHES ON THE DEVELOPMENT AND THE TEST SETS WITH DECISION-FUSION ON ALL FEATURE MODALITIES

	AROUSAL					VALENCE				
MODALITY	SVR+NN	RVM	mvRVM	oaRVM	mvoaRVM	SVR+NN	RVM	mvRVM	oaRVM	mvoaRVM
Dev	.476	.481	.363	**.568**	.481	.461	.406	.343	**.547**	.500
Test	**.444**	-	-	-	.408	.382	-	-	-	**.398**

The results for SVR+NN are obtained from the baseline paper [23]. The test results for RVM, MVRVM, and OARVM are unavailable due to the limitation in the number of results participants are allowed to submit during an active challenge. For each affect dimension and feature modality, the numbers in bold indicate significantly better than the other approaches at $p < .01$ based on the Fisher Z-transform test [30].

consider the same window size for both affect dimensions for simplicity. Alternatively, the window size itself could be learned as a model parameter, as suggested in [14], which is described as part of our future work. Finally, similar to [23], the partially noisy predictions obtained from the RVM-based approaches are smoothed using a median-filter with its window size optimized on the development set using values in the range $[0.2 - 20]s$.

C. Results and Discussion

Due to the AV+EC challenge criteria, the ground truth for the test set is unavailable till date and only a limited number of submissions were allowed to obtain a final decision-fusion score for our proposed approach on the test set. Therefore, all approaches were evaluated on the development set based on both individual feature modalities as well as their decision fusion. But for the test set, the results are only reported for the baseline approaches and our proposed approach on the final decision-fusion outputs.

Results for individual feature modalities on the development set are shown in Table I. In general, audio features perform better for arousal and video features perform better for valence, which is consistent with previous reports [20], [21], [23]. Compared to the audio and video features, both ECG and EDA features perform poorly, possibly because the subjects were moving considerably during the experiments, which could have added noise to the physiological features [21]. A comparison of different approaches for an individual feature modality for each affect dimension shows that, in general, either the OARVM or the proposed MVOARVM significantly ($p < .01$) outperforms the other approaches based on the Fisher Z-transform test [30]. Similarly, results for the decision-fusion in Table II show that on development set, the OARVM significantly ($p < .01$) outperforms the other approaches for predicting both arousal and valence. Furthermore, the best-performing decision-fusion results in Table II significantly ($p < .01$) outperforms the best-performing results using individual feature modalities in Table I on the development set, which shows the benefit of using features from different modalities. On the test set, where the results are only available for the baseline approach and the proposed approach, the baseline approach significantly ($p < .01$) outperforms the proposed approach for predicting arousal and vice versa for predicting valence. In order to analyse the contribution of different feature modalities, the fractions of contributions of different features are plotted in Fig. 2. The values are obtained by normalizing the linear regression weights for each affect dimension obtained with decision-fusion. The figure shows that although individual feature modalities may not perform well by themselves, as evidenced in Table I, in general, all feature modalities contribute towards predicting both affect dimensions, which is consistent with [23]. Moreover, as expected, audio features contribute more towards predicting arousal than valence.

In order to highlight the benefit of the proposed MVOARVM approach compared to the OARVM, the estimated testing time is plotted against the number of affect dimensions to be predicted in Fig. 3. The testing time is estimated rather than computed because the RECOLA database only provides labels for two affect dimensions. Nevertheless, since the testing time for other affect dimensions, if available, are expected to be similar to arousal and valence, the estimates reflect a good approximation of the truth. Fig. 3 shows that while the testing time grows approximately linearly with the number of output variables for the OARVM, it remains more or less constant for the MVOARVM, thus enabling fast affect predictions for the MVOARVM, especially as the number of affect dimensions grows. As described earlier, although arousal and valence are the two most commonly used affect dimensions, other efforts have considered one or more of the additional affect dimensions to model human emotion [15], [16],

[19]. Depending on the application and its requirements, modeling multiple continuous correlated affect dimensions may be important, for which the proposed approach can be implemented without increasing the affect prediction time.

Finally, it is important to also highlight the limitations of the MVOARVM compared to the OARVM and the LSTM-RNN approaches. Training is computationally more expensive for the MVOARVM compared to the OARVM, however both RVM-based approaches are much faster to train compared to the LSTM-RNN [22]. Unlike the OARVM, in the MVOARVM kernel parameters cannot be optimized for each output variable specifically. Instead a single kernel parameter is selected for each regression function modeling all output variables. As a result, given exhaustive parameter optimization, the OARVM is expected to perform better than the MVOARVM. The MVOARVM occasionally outperforms the OARVM in the results shown in Tables I and II only because the kernel parameters were optimized over a coarse grid. In contrast to the LSTM-RNN, where the window size can be learned as part of the model parameter [22], currently, for both RVM-based approaches, the window size must be optimized using methods such as cross-validation. An alternative solution to learning the window size is proposed as part of our future work. Finally, unlike the LSTM-RNN, the RVM-based approaches cannot handle large data, and often, the training data needs to be downsampled. To overcome this limitation, an online learning approach has been proposed as part of our future work.

IV. Conclusions

In this work, motivated by the characteristics of an affect recognition system and inspired by the benefits of the MVRVM and the OARVM, a new dimensional affect prediction approach has been developed that provides fast continuous affect predictions by simultaneously modeling multiple affect dimensions along with their correlations. Our experiments on the RECOLA database has shown that the proposed approach performs competitively with the baseline SVR+NN approach and the OARVM while providing fast predictions than the OARVM during testing.

For future work, the proposed approach can be modified to learn the kernel width and the window size as model parameters, as suggested in [14]. In order to handle large data, the proposed approach can be trained using an online learning framework. Since online learning allows training mini-batches of data, a complete hierarchical Bayesian specification can be used and approximation techniques such as Variational Bayes can be used to perform efficient parameter estimation [28]. Furthermore, in order to learn useful features, the proposed approach can be modified to simultaneously learn the relevance features in addition to the relevance input observations, similar to [32].

Acknowledgment

This work was supported by the U.S. Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate, via a Grant Administered by the Army Research Office under Grant W911NF-09-1-0487 and Grant W911NF-06-1-0357.

References

[1] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huanag, "Human-centred intelligent human? computer interaction (hci^2): how far are we from attaining it?" *International Journal of Autonomous and Adaptive Communications Systems*, vol. 1, no. 2, pp. 168–187, 2008.

[2] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wollmer, "Building autonomous sensitive artificial listeners," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 165–183, April 2012.

[3] M. Mihelj, D. Novak, and M. Munih, "Emotion-aware system for upper extremity rehabilitation," in *Virtual Rehabilitation International Conference, 2009*. IEEE, 2009, pp. 160–165.

[4] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically detecting pain in video through facial action units," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 3, pp. 664–674, 2011.

[5] R. W. Picard, "Future affective technology for autism and emotion communication," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3575–3584, 2009.

[6] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 30–35.

[7] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien, "Emotion on the roadnecessity, acceptance, and feasibility of affective computing in the car," *Advances in human-computer interaction*, vol. 2010, 2010.

[8] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*. Ieee, 2008, pp. 228–235.

[9] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 566–569.

[10] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*. IEEE, 2011, pp. 305–311.

[11] Z. Zeng, M. Pantic, G. Roisman, T. S. Huang *et al.*, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[12] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[13] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 933–936.

[14] "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012.

[15] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[16] R. Dietz and A. Lang, "Affective agents: Effects of agent affect on arousal, attention, liking and learning," in *Proceedings of the Third International Cognitive Technology Conference, San Francisco*, 1999.

[17] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[18] H. Gunes and M. Pantic, "Automatic measurement of affect in dimensional and continuous spaces: Why, what, and how?" in *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*. ACM, 2010, p. 12.

[19] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition*, vol. 17, no. 2, pp. 484–495, 2008.

[20] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14. New York, NY, USA: ACM, 2014, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/2661806.2661807
[21] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, 2014.
[22] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 867–881, 2010.
[23] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "The av+ec 2015 multimodal affect recognition challenge: Bridging across audio, video, and physiological data," 2015.
[24] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. Torr, and R. Cipolla, *Multivariate relevance vector machines for tracking*. Springer, 2006.
[25] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.
[26] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.
[27] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, 10.1007/BF00994018.
[28] C. M. Bishop and M. E. Tipping, "Variational relevance vector machines," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 46–53.
[29] M. E. Tipping, A. C. Faul *et al.*, "Fast marginal likelihood maximisation for sparse bayesian models," in *Proceedings of the ninth international workshop on artificial intelligence and statistics*, vol. 1, no. 3, 2003.
[30] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2007.
[31] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.
[32] B. Krishnapuram, A. Harternink, L. Carin, and M. A. Figueiredo, "A bayesian approach to joint feature selection and classifier design," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1105–1111, 2004.

Achut Manandhar was born in Kathmandu, Nepal. He received the B.S.E.E.(Hons.) degree from Fairleigh Dickinson University, Teaneck, NJ, USA, in 2008. He received the M.S.E.E. degree from Duke University, Durham, NC, USA, in 2012.

He is currently pursuing the Ph.D. degree in electrical engineering at Duke University with research focussed on machine learning approaches for landmine detection and spatio-temporal data analysis.

Kenneth D. Morton Jr. (M'01) was born in York, PA, in 1982. He received the B.S. degree in electrical and computer engineering from The University of Pittsburgh, Pittsburgh, PA, in 2004 and the M.S. and Ph.D. degrees in electrical and computer engineering from Duke University, Durham, NC, in 2006 and 2010 respectively.

He is currently a research scientist at Duke University and Chief Scientist at CoVar Applied Technologies. His research is focused on the development of statistical models for a variety of signal processing and machine learning applications including acoustic signal classification, mapping the sub-surface from ground penetrating radar and understanding video streams.

Dr. Morton is a member of Tau Beta Pi and Eta Kappa Nu.

Peter A. Torrione received the B.S.E.E. degree from Tufts University in 1999, and the Ph.D. degree in electrical engineering from Duke University in 2008, where he became an Assistant Research Professor in 2010.

In 2009 he co-founded New Folder Consulting, which was acquired in 2013 by CoVar Applied Technologies, Inc. He is currently the Chief Technology Officer at CoVar Applied Technologies.

Leslie M. Collins (M'01) was born in Raleigh, NC. She received the B.S.E.E. degree from the University of Kentucky, Lexington, in 1985, and the M.S.E.E. and Ph.D. degrees in electrical engineering, both from the University of Michigan, Ann Arbor, in 1986 and 1995, respectively.

She was a Senior Engineer with the Westinghouse Research and Development Center, Pittsburgh, PA, from 1986 to 1990. She joined Duke in 1995 as an Assistant Professor and was promoted to Associate Professor in 2002. Her current research interests include incorporating physics-based models into statistical signal processing algorithms, and she is pursuing applications in subsurface sensing as well was enhancing speech understanding by hearing impaired individuals.

Dr. Collins is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.