Multimodal Affective Analysis combining Regularized Linear Regression and Boosted Regression Trees

Aleksandar Milchevski[*]
Faculty of Electrical Engineering and Information Technologies
Rugjer Boskovikj 18
Skopje, R. of Macedonia
milchevski@gmail.com

Alessandro Rozza
HYERA Software
Via Mattei 2
Coccaglio, Italy
alessandro.rozza@hyera.com

Dimitar Taskovski
Faculty of Electrical Engineering and Information Technologies
Rugjer Boskovikj 18
Skopje, R. of Macedonia
dtaskov@feit.ukim.edu.mk

ABSTRACT

In this paper we present a multimodal approach for affective analysis that exploits features from video, audio, Electrocardiogram (ECG), and Electrodermal Activity (EDA) combining two regression techniques, namely Boosted Regression Trees and Linear Regression. Moreover, we propose a novel regularization approach for the Linear Regression in order to exploit the temporal correlation of the affective dimensions. The final prediction is obtained using a decision level fusion of the regressors individually trained on the different groups of features. The promising results obtained on the benchmark dataset show the efficacy and effectiveness of the proposed approach.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern recognition—*Applications*

Keywords

Affective Computing, Linear Regression, Regression Trees, Regularization

1. INTRODUCTION

In the last decade, affective computing has grown its importance becoming a hot research topic. The emotion recognition guarantees many benefits since our emotions play a fundamental role in everyday communications. The aim of an affective computing approach is to recognize the emotional state of the analyzed target. Usually, face imagery (or videos) are used to analyze the facial expressions of the target as clue for the affective state. Moreover, the speech of the analyzed target is also used to infer the affective state. An efficient affective computing approach exploits this information to make the final prediction. Due to the complex nature of emotions, a multimodal approach (an approach which combines several information sources) is better suited to achieve good assessment of the affective state.

In this paper we propose a novel multimodal affective approach that combines two different regression techniques. Moreover, a novel regularization approach for regression is proposed to exploit the temporal correlation of the affective dimensions improving the results. The paper is organized as follows: in Sec. 2 a brief overview of state-of-the-art is provided; in Sec. 3 we describe each building block of our approach; in Sec. 4 the experimental results are presented; finally, conclusions are summarized in Sec. 5.

2. RELATED WORK

In this section we briefly describe some of the most important works in this field.

In contrast with traditional class-based emotion recognition approaches, in their seminal work the authors of [27] introduce the continuous emotion recognition by employing Conditional Random Field (CRF, [8]) and Long Short Term Memory Recurrent Neural Networks (LSTM-RNN, [6]). This approach clearly shows the advantages of using temporal information.

In [13] the authors propose to employ Latent Dynamic Conditional Random Fields (LDCRF, [11]). This method exploits the sub-structure of the affective signals as well as the extrinsic dynamic between emotional labels. In this paper the best results are obtained by using the LDCRF to fuse the outputs of the unimodal classifiers. Similarly, the approach proposed in [20] combines the information from audio and video features for emotion classification. Moreover, Partial Least Squares Regression (PLSR, [19]) is used to reduce the feature dimensionality. Subsequently, the kernel SVM outputs are used as inputs of a graphical model (Joint Hidden Conditional Random Field, JHCRF). The JHCRF fuses the decisions of the classifiers exploiting the temporal correlation between the labels to improve the results.

In [1] the authors use a hybrid model applying Conditional Restricted Bolzman Machine (CRBM, [21]) and CRF. According to the authors, the deep networks based temporal generative model captures the short term temporal characteristics, and the CRF captures the long range temporal de-

[*] The author is a PH.D. student at the Faculty of Electrical Engineering and Information Technologies and is also affiliated to Nagi Korporacija-Skopje

AVEC'15, October 26 2015, Brisbane, Australia
©2015 ACM. ISBN 978-1-4503-3743-4/15/10...$15.00.
DOI: http://dx.doi.org/10.1145/2808196.2811636.

pendencies. In [3] a similar idea is applied for dimensional affective analysis. The authors combine Continuous Conditional Random Fields (CCRF, [12]) and Support Vector Machines for Regression to model continuous emotions.

In [14, 16] the authors predict dimensional emotion ratings and introduce the RECOLA dataset. This dataset is composed by audio-visual and physiological data.

Another framework useful to combine information from different groups of features is the Multiple Kernel Learning (MKL, [2]). In [7] the authors use a MKL based approach in order to select the best audio features and to perform the final classification. MKL is also used in [9] for emotion recognition in the wild challenge, where this methodology achieved the best result[5].

Finally, it is important to recall that the most widely used benchmarks to evaluate the affective computing approaches are those proposed by the Audio Visual Emotional Challenges (AVEC) [17, 18, 24, 23] organized since 2011.

3. OUR APPROACH

In this section, we describe the methodologies behind the proposed multimodal approach for affective analysis. These methods are separately summarized in Sec. 3 while the description of the best combination of these techniques is presented in Sec. 4. We present the two regression techniques employed and a novel regularization that exploits temporal correlation between the labels. Furthermore, we describe the explicit kernel feature mapping and the simple smoothening filter that are used to improve the predictive results. Finally, four different decision fusion schemes and a feature level fusion schema are introduced. Our system outputs a continuous prediction for two affective dimensions: Arousal and Valence.

3.1 Boosted Regression Trees

The first regression technique employed in our tests is the Boosted Regression Trees [10]. This approach is an ensemble method that combines the strengths of two algorithms: regression trees, a technique that relates a response to their predictors by recursive binary splits, and boosting, an adaptive method for combining many simple models to improve predictive performance. We employ a regression tree with a minimal of 75 observation per tree leaf. Fifteen regression trees are trained using a gradient boosting approach with a learning rate of 0.2. The same method is employed for feature selection (see Sec. 4 for more details).

3.2 A Novel Regularized Linear Regression

Linear regression performs a simple scalar product between the feature vectors and the parameters computed during the training phase. The training consists of finding the parameters that best fit the training data (i.e. minimize the cost function).

To improve our results, we propose a novel regularization that exploits the temporal correlation between labels. To better understand this novel regularization the following cost function is introduced:

$$e^{(i)} = (\mathbf{x}^{(i)}\boldsymbol{\theta} - y^{(i)})^2 + \lambda_2 \sum_{k=-N_g}^{N_g} \mathcal{N}(k)(\mathbf{x}^{(i)}\boldsymbol{\theta} - \mathbf{x}^{(i+k)}\boldsymbol{\theta})^2 \quad (1)$$

$$J(\boldsymbol{\theta}) = \frac{1}{2m}[\sum_{i=1}^{m} e^{(i)} + \lambda_1||\tilde{\boldsymbol{\theta}}||^2], \quad (2)$$

where m is the total number of feature vectors, $\boldsymbol{\theta}$ is a vector with all the parameters obtained during the training phase, N_g is the size of the neighborhood for the regularization, $\mathbf{x}^{(i)}$ and $y^{(i)}$ are respectively the i-th feature vector and the corresponding label. The first element in $\mathbf{x}$ is always equal to 1 to account the bias term and make the equations simpler. Furthermore, $\tilde{\boldsymbol{\theta}}$ is a vector of parameters and $\mathcal{N}(k)$ is a Gaussian function.

The first term in Eq. 1 is the squared error between the prediction and the actual label while the second is the first regularization term. This regularization term penalizes the difference between neighboring temporal predictions. The term represents the sum of the squares of the differences between neighboring predictions in a given neighborhood weighted by a Gaussian function, which reduces the effect of the penalization when the time difference between the samples increases. Assuming that the value for the affective dimension does not change dramatically in short time periods, embedding this term in the error function will lead to more stable and better solutions. Eq. 2 gives the cost function that is minimized in order to find the theta parameters. The second term of Eq. 2 represents the second type of regularization that is employed. The term represents the Tikhonov regularization [22], which is commonly used to simplify the solutions (i.e. prevent overfitting on the training set).

In order to obtain the gradient, the cost function is rewritten in the following matrix form:

$$J(\boldsymbol{\theta}) = \frac{1}{2m}[(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) + \lambda_1||\tilde{\boldsymbol{\theta}}||^2] + \lambda_2 R(\boldsymbol{\theta}) \quad (3)$$

$$R(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{k=-Ng}^{Ng} \mathcal{N}(k)(\mathbf{X}\boldsymbol{\theta} - \mathbf{X}_k\boldsymbol{\theta})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{X}_k\boldsymbol{\theta}), \quad (4)$$

where $\mathbf{y}$ is a vector of all the labels, $\mathbf{X} = [\mathbf{x}^{(1)T}; ...\mathbf{x}^{(m)T}]$ is a matrix of all of the feature vectors, and $\mathbf{X}_k$ is the matrix with all the feature vectors shifted of k positions. In order to simplify the equations, the matrices $\mathbf{X}_k$ are of the same size.

For positive values of k the matrix is the following:

$$\mathbf{X}_k = [\mathbf{x}^{(1)T}; \mathbf{x}^{(2)T}; ...\mathbf{x}^{(k)T}; \mathbf{x}^{(1)T}; \mathbf{x}^{(2)T}...\mathbf{x}^{(m-k)T}] \quad (5)$$

For negative values of k the matrix is similar except for the last k rows of $\mathbf{X}_k$ and $\mathbf{X}$ that are set to be equal.

Solving the gradient of the cost function, we obtain the following equations:

$$\nabla_\theta J(\boldsymbol{\theta}) = \frac{1}{m}[\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y} + \lambda_1\tilde{\mathbf{I}}\boldsymbol{\theta}] + \lambda_2\nabla_\theta R(\boldsymbol{\theta}) \quad (6)$$

$$\nabla_\theta R(\boldsymbol{\theta}) = \frac{1}{m} \sum_{k=-Ng}^{Ng} \mathcal{N}(k)(\mathbf{X} - \mathbf{X}_k)^T(\mathbf{X} - \mathbf{X}_k)\boldsymbol{\theta} \quad (7)$$

where $\tilde{\mathbf{I}}$ is the identity matrix with the first element of the diagonal set to zero in order to avoid the regularization of the bias term.

By employing the formalized cost function and its gradient, a generic solver can be used to minimize the cost

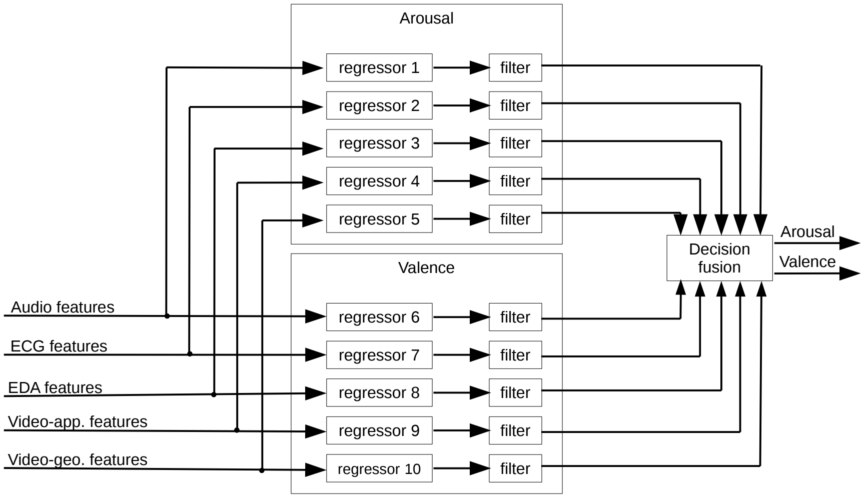

Figure 1: General block diagram of the proposed approaches using decision fusion.

function and to obtain the theta parameters. We employ the fminunc function from MATLAB with the trust-region-reflective algorithm [4].

3.3 Explicit Feature Mapping

In order to overcome the limitation of linear methods imposed by the employed regression techniques we propose to use an explicit feature mapping. Vedaldi et al. [26] have shown that the approximate explicit feature mapping achieves indistinguishable performance with respect to those obtained by a full intersection kernel, while greatly reducing the time costs during the training and the classification phases.

To compute the explicit feature mapping we have employed the Vlfeat open source library [25] using a χ^2-kernel and tuning the kernel parameter.

3.4 Filtering

With the aim of improving our results we employ a simple smoothening filter. This approach guarantees to obtain smooth predictions that usually describe better the transitions between the different affective states. This filter is formalized as follows:

$$y_f^{(i)} = (1 - k_g)y_f^{(i-1)} + k_g y^{(i)}, \qquad (8)$$

where $y_f^{(i)}$ is i-th filtered prediction and $y^{(i)}$ is the unfiltered prediction label by the regression technique. The k_g coefficient is constant and has been empirically set to 0.03.

3.5 Information Fusion

We have evaluated two fusion schemas: the first one works on the feature level and the second on the decision level. The feature level fusion simply concatenates the different feature groups in a single feature vector. The decision level fusion is done jointly for arousal and valence using linear regression that combines different models learned on the training set (for more details see Sec. 4).

4. EXPERIMENTAL RESULTS

For the evaluation of our method the Audio/Visual$^+$ Emotional Challenge (AV$^+$EC) 2015 dataset is used. Precisely, we use five sets of features extracted from video, audio, Electrocardiogram (ECG), and Electrodermal Activity (EDA) signals (see Sec. 4.1). The test set is not available to the authors and the evaluation for this set is done by the challenge organizers. The employed performance metric is the Concordance Correlation Coefficient (CCC), which combines the Pearson's correlation coefficient (CC) with the mean-square error.

In order to evaluate the quality of our approach and to produce its final configuration we have performed different experiments. The first set of experiments is related to the audio feature extraction and aimed to analyze the different performance of the employed regressors changing the size of the time windows (see Sec. 4.2). The second set of experiments (see Sec. 4.3) shows the performance achieved by our regression techniques trained only on a single feature group (video, audio, ECG, or EDA). In the third set of experiments (see Sec. 4.4) the regression techniques trained in the previous experiments are combined using different approaches so to obtain the final estimation results.

4.1 Features

We propose a multimodal approach for affective analysis that uses five groups of features. Two groups of features are extracted from the video and three groups are respectively extracted from audio, ECG, and EDA signals.

For the video, ECG, and EDA features we use the baseline features proposed in the AV$^+$EC 2015. For a detailed description of these features the reader can refer to the baseline papers [15] published by the organizers of the challenge.

For the audio features we use those proposed for the AVEC 2011 [17] challenge varying the time window size. The audio feature extraction follows a brute force approach. A large number of features are extracted from each time window of the analyzed audio signal. In Tab. 1 the low level descriptors (LLD) extracted from each time window are presented. On the LLD the functionals presented in Tab. 2 are applied

Table 1: Low Level Descriptors.

Energy & Spectral (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250-650 Hz, 1kHz-4kHz, 25%, 50%, 75%, and 90% spectral roll-off points, spectral flux, entropy, spectral variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity, MFCC 1-10
Voicing related (6)
F0 (Sub-harmonic summation (SHS) followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), delta: jitter of jitter, logarithmic Harmonics-to-Noise Ratio (logHNR)

Table 2: Functionals.

Statistical functionals (23)
arithmetic mean, root quadratic mean standard deviation, flatness skewness, kurtosis quartiles, and inter-quartile ranges 1%, 99% percentile percentile range 1%-99% percentage of frames contour is above: (min + 25%, 50%, and 90%) of the range percentage of frames contour is rising max, mean, min segment length standard deviation of segment length logarithmic Harmonics-to-Noise Ratio (logHNR)
Regression functionals (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient and approximation error (linear)
Local minima/maxima related functionals (9)
mean and standard deviation of rising and falling slopes, mean and standard deviation of inter maxima distances amplitude mean of maxima amplitude mean of minima amplitude range of maxima
Other (6)
LP gain, LPC 1-5

obtaining the final feature vector with a dimensionality of 1941.

4.2 Time Window Analysis

In Tab. 3 the results obtained changing the size of the time windows during the audio feature extraction process are summarized. All the feature vectors are created by computing the audio features and concatenating them with the precomputed video, ECG, and EDA features. We have tested both Boosted Regression Trees and Regularized Linear Regression. The outputs from both regression techniques are filtered using the previously described smoothening filter. Similarly to the AV$^+$EC 2015 baseline paper, a subsample of training dataset is employed during the training phase. Precisely, only the 10-th feature vectors and their corresponding labels are used to create the training set. The algorithm is evaluated on the whole development set.

Table 3: Results (CCC) obtained using different window sizes for the audio feature extraction.

Window size	1s	3s	5s	7s
Arousal - lin. reg.	0.3160	0.5573	0.6351	**0.6653**
Valence - lin. reg.	0.2131	**0.2386**	0.2358	0.206
Arousal - b.r.trees	0.3623	0.4945	0.5488	**0.5872**
Valence - b.r.trees	**0.3026**	0.2814	0.2568	0.2457

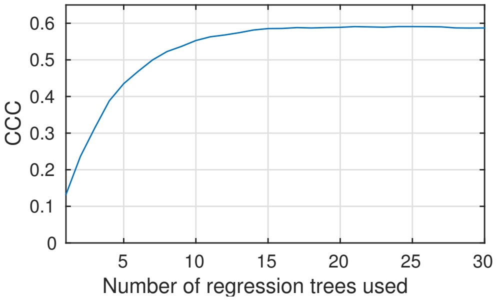

Figure 2: The CCC as a function of the number of the regression trees used.

The results presented in Tab. 3 show that the size of the time window strongly affects the accuracy of the affective analysis algorithm. Moreover, these results show that using large time windows for the audio feature extraction achieve better results for Arousal, instead small time windows guarantee the best results for Valence.

In Fig. 2 we show the different results of the Boosted regression trees varying the number of the employed trees (from 1 to 30). In particular, the CCC is shown for the predictions computed on the development set depending on the number of trees employed. These results clearly show the benefit of boosting the regression trees and that the CCC does not significantly increase after 15 trees.

4.3 Tests on the different feature groups

In these experiments we have trained both the regularized linear regression and the boosted regression trees on each feature group. Considering the regularized linear regression a pre-processing step to reduce the feature dimensionality is employed. Precisely, we have computed the feature importance values by using the boosted regression trees method and we have selected all the features with an importance value greater than a fixed threshold. The outputs of all of the regressors are filtered with the previously described smoothening filter. The approaches are trained and evaluated on the same training and development sets described in the previous section.

The results summarized in Tab. 4 show that the proposed regularized linear regression method strongly overcomes the baseline method on the audio features. Moreover, the boosted regression trees perform better than regularized linear regression on the video features and obtain comparable results with those achieved by the baseline approach proposed by the challenge organizers.

Table 4: Results (CCC) obtained using the different feature groups.

Method	Aff. dimension	Audio	ECG	EDA	Video - appearance	Video - geometric
Baseline	Arousal	0.287	**0.275**	0.078	0.103	**0.231**
Boosted regression trees	Arousal	0.572	0.180	0.056	**0.112**	0.056
Regularized linear regression	Arousal	**0.649**	0.166	**0.090**	0.023	0.073
Baseline	Valence	0.069	**0.183**	**0.204**	**0.273**	**0.325**
Boosted regression trees	Valence	0.100	0.120	0.107	0.196	0.259
Regularized linear regression	Valence	**0.120**	0.152	0.113	0.093	0.237

4.4 Decision Level Fusion

In these experiments we have analyzed several approaches to combine the regressors described in the previous section to improve the final results. The general schema for the decision level fusion is summarized in Fig. 1.

The first fusion approach combines the individual linear regularized regressors trained on the different feature groups. Furthermore, for valence, we have applied the explicit feature mapping described in Sec. 3.3. The decision fusion is jointly performed for both the affective dimensions using a linear regression approach.

The results summarized in Tab. 5 and Tab. 6 show that the CCC is higher for both the development and test set compared with those obtained by the baseline method proposed by the organizers.

The second fusion approach combines the regressors that achieves better performance on the development set (for example, for the regressor predicting the arousal dimension trained with the audio features we have used the regularized linear regression approach and for the regressors predicting the valence dimension trained with the video features we have used the boosted regression trees).

The results reported in Tab. 5 and Tab. 6 show that this approach achieves better performance compared to those obtained by the baseline approach and also compared to those obtained by the first fusion schema.

The third fusion approach combines the regressors trained with the proposed regularized linear approach. The difference from the first fusion is that explicit kernel mapping is employed also on the Arousal features.

Compared with the previous two fusion approaches this methodology yields better performance for Arousal dimension on both the development and test set, however for Valence dimension the performance is worse.

Considering the fourth fusion approach we have used two regressors for every feature group, one with the regularized linear regression and the other with the boosted regression trees. The obtained results show an improvement for Arousal, but the performance for Valence is the lowest.

Considering all the results obtained, the best prediction model combines the results achieved by the second fusion schema (for valence) and those obtained by the last fusion schema (for arousal). In Fig. 3 we summarize this prediction model. The second schema is particularly suitable for valence reducing the overfitting on the test set, while the fourth fusion schema obtains good results for arousal but overfits on the valence. Finally, it is important to underline that we have performed other tests (not reported in tables for reasons of clarity) where we have used separate regressors for each dimension. From these tests we have noticed that the results are always worse thus inducing to think that there are some useful relations that can be exploited predicting jointly the two dimensions.

Table 5: Final results (CCC) - Arousal.

Set	Development	Test
Baseline	0.476	0.444
Fusion No1	0.756	0.504
Fusion No2	0.755	0.508
Fusion No3	0.758	0.633
Fusion No4	**0.791**	**0.644**

Table 6: Final results (CCC) - Valence.

Set	Development	Test
Baseline	0.461	0.382
Fusion No1	0.495	0.473
Fusion No2	0.516	**0.501**
Fusion No3	0.485	0.345
Fusion No4	**0.542**	0.251

5. CONCLUSIONS

In this paper we have presented a multimodal affective analysis approach that combines regularized linear regression and boosted regression trees. A novel regularization technique has been proposed in order to use the temporal correlation of the labels of the affective dimensions. Experiments performed on the AV$^+$EC 2015 dataset has confirmed the quality and the effectiveness of the proposed approach. Moreover, an important contribution of the proposed methodology consists in its better computational efficiency compared with many state-of-the-art approaches.

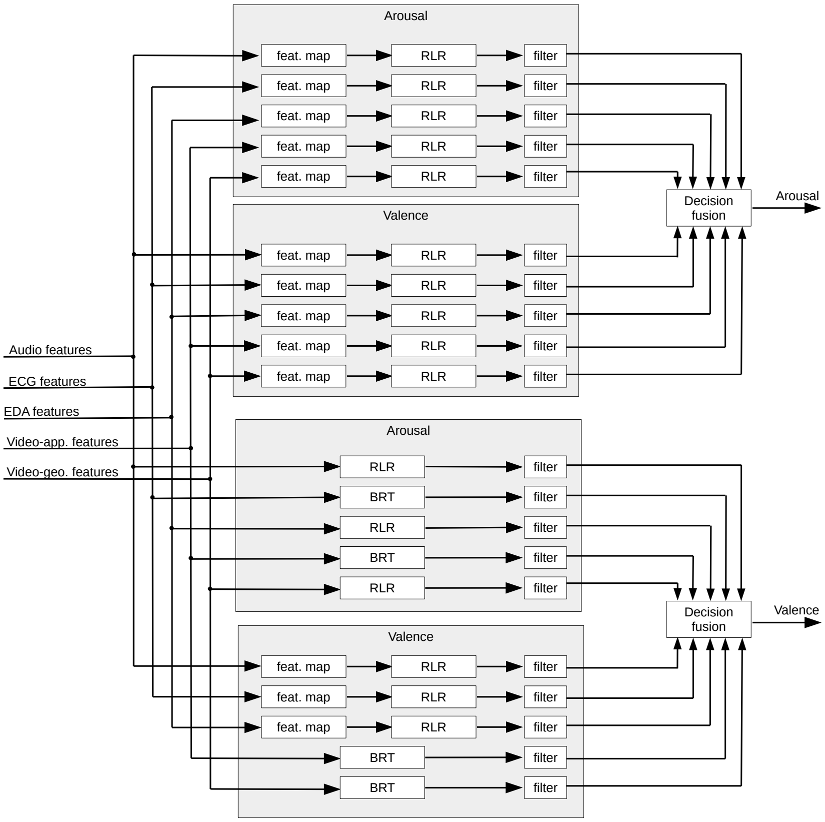

Figure 3: The best performing approach.

6. REFERENCES

[1] M. R. Amer, B. Siddiquie, C. Richey, and A. Divakaran. Emotion detection in speech using deep networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3724–3728. IEEE, 2014.

[2] F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

[3] T. Baltrusaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[4] T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization*, 6(2):418–445, 1996.

[5] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 461–466. ACM, 2014.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Y. Jin, P. Song, W. Zheng, and L. Zhao. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4808–4812. IEEE, 2014.

[8] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[9] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 494–501. ACM, 2014.

[10] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent in function space. NIPS, 1999.

[11] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[12] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li. Global ranking using continuous conditional random fields. In *Advances in neural information processing systems*, pages 1281–1288, 2009.

[13] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Affective Computing and Intelligent Interaction*, pages 396–406. Springer, 2011.

[14] 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.

[15] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC), ACM MM*, Brisbane, Australia, October 2015.

[16] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proceedings of Face & Gestures 2013, 2nd IEEE International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Shanghai, China, April 2013.

[17] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

[18] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[19] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *Computer vision, 2009 IEEE 12th international conference on*, pages 24–31. IEEE, 2009.

[20] B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney. Affect analysis in natural human interaction using joint hidden conditional random fields. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[21] G. W. Taylor, G. E. Hinton, and S. T. Roweis. Modeling human motion using binary latent variables. In *Advances in neural information processing systems*, pages 1345–1352, 2006.

[22] A. N. Tikhonov and V. Arsenin. *Solutions of ill-posed problems*. Vh Winston, 1977.

[23] 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*, pages 3–10. ACM, 2014.

[24] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[25] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, pages 1469–1472. ACM, 2010.

[26] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492, 2012.

[27] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes – towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, volume 2008, pages 597–600, 2008.