HAL
open science

Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals

Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, Björn Schuller

To cite this version:

Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, Björn Schuller. Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals. Image and Vision Computing, 2017, 65, pp.76-86. 10.1016/j.imavis.2016.11.020 . hal-01486186

HAL Id: hal-01486186

https://hal.science/hal-01486186v1

Submitted on 9 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals

Jing Han[a], Zixing Zhang[a,*], Nicholas Cummins[a], Fabien Ringeval[a,b], Björn Schuller[a,c]

[a]*Chair of Complex and Intelligent Systems, University of Passau, Innstr. 41, 94032 Passau, Germany*
[b]*Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, 700 Avenue Centrale, 38058 Grenoble, France*
[c]*Department of Computing, Imperial College London, 180 Queens' Gate, London SW7 2AZ, UK*

Abstract

Automatic continuous affect recognition from audiovisual cues is arguably one of the most active research areas in machine learning. In addressing this regression problem, the advantages of the models, such as the global-optimisation capability of Support Vector Machine for Regression and the context-sensitive capability of memory-enhanced neural networks, have been frequently explored, but in an isolated way. Motivated to leverage the individual advantages of these techniques, this paper proposes and explores a novel framework, Strength Modelling, where two models are concatenated in a hierarchical framework. In doing this, the strength information of the first model, as represented by its predictions, is joined with the original features, and this expanded feature space is then utilised as the input by the successive model. A major advantage of Strength Modelling, besides its ability to hierarchically explore the strength of different machine learning algorithms, is that it can work together with the conventional feature- and decision-level fusion strategies for multimodal affect recognition. To highlight the effectiveness and robustness of the proposed approach, extensive experiments have been carried out on two time- and value-continuous spontaneous emotion databases (RECOLA and SEMAINE) using audio and video signals. The experimental results indicate that employing Strength Modelling can deliver a significant performance improvement for both arousal and valence in the unimodal and bimodal settings. The results further show that the proposed systems is competitive or outperform the other state-of-the-art approaches, but being with a simple implementation.

1. Introduction

Automatic affect recognition plays an essential role in smart conversational agent systems that aim to enable natural, intuitive, and friendly human–machine interaction. Early works in this field have focused on the recognition of prototypic expressions in terms of basic emotional states, and on the data collected in laboratory settings, where speakers either act or are induced with predefined emotional categories and content [9, 29, 30, 47]. Recently, an increasing amount of research efforts have converged into dimensional approaches for rating naturalistic affective behaviours by continuous dimensions (e. g., arousal and valence) along the time continuum from audio, video, and music signals [8, 10, 24, 39, 46, 16, 32, 33]. This trend is partially due to the benefits of being able to encode small difference in affect over time and distinguish the subtle and complex spontaneous affective states. Furthermore, the affective computing community is moving toward combining multiple modalities (e. g., audio and video) for the analysis and recognition of human emotion [19, 23, 34, 43, 49], owing to (i) the easy access to various sensors like camera and microphone, and (ii) the complementary information that can be given from different modalities.

In this regard, this paper focuses on the realistic time- and value-continuous affect (emotion) recognition from audiovisual signals in the arousal and valence dimensional space. To handle this regression task, a variety of models have been investigated. For instance, *Support Vector Machine for Regression* (SVR) is arguably the most frequently employed approach owing to its mature theoretical foundation. Further, SVR is regarded as a baseline regression approach for many continuous affective computing tasks [27, 31, 36]. More recently, memory-enhanced *Recurrent Neural Networks* (RNNs), namely *Long Short-Term Memory RNNs* (LSTM-RNNs) [14], have started to receive greater attention in the sequential pattern recognition community [7, 26, 48, 50]. A particular advantage offered by LSTM-RNNs is a powerful capability to learn longer-term contextual information through the implementation of three memory gates in the hidden neurons. Wöllmer et al. [41] was amongst the first to apply LSTM-RNN on acoustic features for continuous affect recognition. This technique has also been successfully employed for other modalities (e. g., video, and physiological signals) [2, 21, 26].

Numerous studies have been performed to compare the advantages offered by a wide range of modelling techniques, including the aforementioned, for continuous affect recognition [21, 27, 35]. However, no clear observations can be drawn as to the superiority of any of them. For instance, the work in [21] compared the performance of SVR and *Bidirec-*

*corresponding author: zixing.zhang@uni-passau.de, Tel.: +49 851 509-3359, Fax.: +49 851 509-3352

tional LSTM-RNNs (BLSTM-RNNs) on the Sensitive Artificial Listener database [20], and the results indicate that the latter performed better on a reduced set of 15 acoustic *Low-Level-Descriptors* (LLD). However, the opposite conclusion was drawn in [35], where SVR was shown to be superior to LSTM-RNNs on the same database with functionals computed over a large ensemble of LLDs. Other results in the literature confirm this inconsistent performance observation between SVR and diverse neural networks like (B)LSTM-RNNs and *Feed-forward Neural Networks* (FNNs) [27]. A possible rationale behind this is the fact that each prediction model has its advantages and disadvantages. For example, SVRs cannot explicitly model contextual dependencies, whereas LSTM-RNNs are highly sensitive to overfitting.

The majority of previous studies have tended to explore the advantages (strength) of these models independently or in conventional early or late fusion strategies. However, recent results indicate that there may be significant benefits in fusing two, or more, models in hierarchical or ordered manner [15, 18, 22]. Motivated by these initial promising results, we propose a *Strength Modelling* approach, in which the strength of one model, as represented by its predictions, is concatenated with the original feature space which is then used as the basis for regression analysis in a subsequent model.

The major contributions of this study include: (1) proposing the novel machine learning framework of Strength Modelling specifically designed to take advantage of the benefits offered by various regression models namely SVR and LSTM-RNNs; (2) investigating the effectiveness of Strength Modelling for value- and time-continuous emotion regression on two *spontaneous multimodal* affective databases (RECOLA and SEMAINE); and (3) comprehensively analysing the robustness of Strength Modelling by integrating the proposed framework into frequently used multimodal fusion techniques namely early and late fusion.

The remainder of the present article is organised as follows: Section 2 first discusses related works; Section 3 then presents Strength Modelling in details and briefly reviews both the SVR and memory-enhanced RNNs; Section 4 describes the selected spontaneous affective multimodal databases and corresponding audio and video feature sets; Section 5 offers an extensive set of experiments conducted to exemplify the effectiveness and the robustness of our proposed approach; finally, Section 6 concludes this work and discusses potential avenues for future work.

2. Related Work

In the literature for multimodal affect recognition, a number of fusion approaches have been proposed and studied [45], with the majority of them relevant to *early* (aka *feature-level*) or *late* (aka *decision-level*) fusion. Early fusion is implemented by concatenating all the features from multiple modalities into one combined feature vector, which will then be used as the input for a machine learning technique. The benefit of early fusion is that, it allows a classifier to take advantage of the complementarity that exists between, for example, the audio and video feature spaces. The empirical experiments offered in [2, 15, 26] have shown that the early fusion strategy can deliver better results than the strategies without feature fusion.

Late fusion involves combining predictions obtained from individual learners (models) to come up with a final prediction. They normally consist of two steps: 1) generating different learners; and 2) combining the predictions of multiple learners. To generate different learners, there are two primary ways which are separately based on different *modalities* and *models*. Modality-based ways combines the output from learners trained on *different* modalities. Examples of this learner generation in the literature include [12, 15, 22, 37], where multiple SVRs or LSTM-RNNs are trained separately for different modalities (e.g. audio, video, etc). Model-based ways, on the other hand, aims to exploit information gained from multiple learners trained on a single modality. For example in [25], predictions obtained by 20 different topology structures of Deep Belief Networks (DBNs). However, due to the similarity of characteristics of different DBNs, the predictions can not provide many variations that could be mutually complemented and improve the system performance. To combine the predictions of multiple learners, a straightforward way is to apply simple or weighted averaging (or voting) approach, such as Simple Linear Regression (SLR) [36, 15]. Another common approach is to perform stacking [44]. In doing this, all the predictions from different learners are stacked and used as inputs of a subsequent non-linear model (e.g., SVR, LSTM-RNN) trained to make a final decision [25, 12, 37].

Different from these fusion strategies, our proposed Strength Modelling paradigm operates on a *single* feature space. Using an initial model, it gains a set of predictions which are then fused with the original feature set for use as a new feature space in a subsequent model. This offers the framework a vital important advantage as the single modality setting is often faced in affect recognition tasks, for example, if when either face or voice samples are missing in a particular recording.

Indeed, Strength Modelling can be viewed as an *intermediate* fusion technology, which lies in the middle of the early and late fusion stages. Strength Modelling can therefore not only work independently of, but also be simply integrated into early and late fusion approaches. To the best of our knowledge, intermediate fusion techniques are not widely used in the machine learning community. Hermansky et al. [13] introduced a tandem structure that combines the output of a discriminative trained neural nets with dynamic classifiers such as *Hidden Markov Models* (HMMs), and applied it efficiently for speech recognition. This structure was further extended into a BLSTM-HMM [40, 42]. In this approach the BLSTM networks provides a discrete phoneme prediction feature, together with continuous *Mel-Frequency Cepstral Coefficients* (MFCCs), for the HMMs that recognise speech.

For multimodal affect recognition, a relevant approach – Parallel Interacting Multiview Learning (PIML) – was proposed in [17] for the prediction of protein sub-nuclear locations. The approach exploits different modalities that are mutually learned in a parallel and hierarchical way to make a final decision. Reported results show that this approach is more suitable than the

use of early fusion (merging all features). Compared to our approach, that aims at taking advantages of different models from a same modality, the focus of PIML is rather on exploiting the benefit from different modalities. Further, similar to early fusion approaches, PIML operates under a concurrence assumption of multiple modalities.

Strength Modelling is similar to the *Output Associative Relevance Vector Machine* (OA-RVM) regression framework originally proposed in [22]. The OA-RVM framework attempts to incorporate the contextual relationships that exist within and between different affective dimensions and various multimodal feature spaces, by training a secondary RVM with an initial set of multi-dimensional output predictions (learnt using any prediction scheme) concatenated with the original input features spaces. Additionally, the OA-RVM framework also attempts to capture the temporal dynamics by employing a sliding window framework that incorporates both past and future initial outputs into the new feature space. Results presented in [15] indicate that the OA-RVM framework, is better suited to affect recognition problems than both conventional early and late fusion. Recently the OA-RVM model was extended in [18] to be multivariate, i. e., predicting multiple continuous output variables simultaneously.

Similar to Strength Modelling, OA-RVM systems take input features and output predictions into consideration to train a subsequent regression model to perform the final affective predictions. However, the strength of the OA-RVM framework is that it is underpinned by the RVM. Results in [15] indicate that, the framework is not as successful when using either a SVR or a SLR as the secondary model. Further, the OA-RVM is non-casual and requires careful tuning to find suitable window lengths in which to combine the initial outputs; this can take considerable time and effort. The proposed Strength Modelling framework, however, is designed to work with *any* combination of learning paradigms. Furthermore, Strength Modelling is casual; it combines input features and predictions on a frame-by-frame basis. This is a strong advantage over the OA-RVM in terms of employment in real-time scenarios (beyond the scope of this paper).

3. Strength Modelling

3.1. Strength Modelling

The proposed Strength Modelling framework for affect prediction is depicted in Fig. 1. As can be seen, the first regression model ($Model_1$) generates the original estimate $\hat{y}_t$ based on the feature vector $\boldsymbol{x}_t$. Then, $\hat{y}_t$ is concatenated with $\boldsymbol{x}_t$ pair-wise as the input of the second model ($Model_2$) to learn the expected prediction y_t.

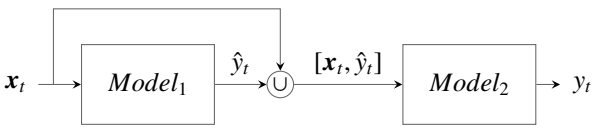


Figure 1: Overview of the Strength Modelling framework.

To implement the Strength Modelling for these suitable combination of individual models, $Model_1$ and $Model_2$ are trained subsequently, in other words, $Model_2$ takes the predictive ability of $Model_1$ into account for training. The procedure is given as follows:

- First, $Model_1$ is trained with $\boldsymbol{x}_t$ to obtain the prediction $\hat{y}_t$.
- Then, $Model_2$ is trained with $[\boldsymbol{x}_t, \hat{y}_t]$ to learn the expected prediction y_t.

Whilst the framework should work with any arbitrary modelling technique we have selected two commonly used, in the context of affect recognition, for our initial investigations, namely the SVR and BLSTM-RNNs which are briefly reviewed in the subsequent subsection.

3.2. Regression Models

SVR is extended from Support Vector Machine (SVM) to solve regression problems. It was first introduced in [4] and is one of the most dominant methods in the context of machine learning, particularly in emotion recognition [1, 27]. Applying the SVR for a regression task, the target is to optimise the generalisation bounds for regression in the high-dimension feature space by using a ε-insensitive loss function which is used to measure the cost of the errors of the prediction. At the same time, a predefined hyperparameter C is set accordingly for different cases to balance the emphasis on the errors and the generalisation performance.

Normally, the high-dimension feature space is mapped from the initial feature space with a non-linear kernel function. However, in our study, we use a linear kernel function, as the features in our cases (cf. Section 4.2) perform quite well for affect prediction in the original feature space, similar to [36].

One of the most important advantages of SVR is the convex optimisation function, the characteristics of which gives the benefit that the global optimal solution can be obtained. Moreover, SVR is learned by minimising an upper bound on the expected risk, as opposed to the neural networks trained by minimising the errors on all training data, which equips SVR a superior ability to generalise [11]. For a more in-depth explanation of the SVR paradigm the reader is referred to [4].

The other model utilised in our study is BLSTM-RNN which has been successfully applied to continuous emotion prediction [26] as well as for other regression tasks, such as speech dereverberation [48] and non-linguistic vocalisations classification [24]. In general, it is composed of one input layer, one or multiple hidden layers, and one output layer [14]. The bidirectional hidden layers separately process the input sequences in a forward and a backward order and connect to the same output layer which fuses them.

Compared with traditional RNNs, it introduces recurrently connected memory blocks to replace the network neurons in the hidden layers. Each block consists of a self-connected memory cell and three gate units, namely input, output, and forget gate. These three gates allow the network to learn when to write, read, or reset the value in the memory cell. Such a structure grants BLSTM-RNN to learn past and future context

in both short and long range. For a more in-depth explanation of BLSTM-RNNs the reader is referred to [14].

It is worth noting that these paradigms bring distinct sets of advantages and disadvantages to the framework:

- The SVR model is more likely to achieve the global optimal solution, but it is not context-sensitive [21];
- The BLSTM-RNN model is easily trapped in a local minimum which can be hardly avoided and has a risk of overfitting [7], while it is good at capturing the correlation between the past and the future information [21].

In this paper, $Model_1$ and $Model_2$ in Fig. 1 could be either an SVR model or a BLSTM-RNN model, resulting in four possible permutations, i. e., SVR-SVR (S-S), SVR-BLSTM (S-B), BLSTM-SVR (B-S), BLSTM-BLSTM (B-B). It is worth noting that the B-B structure can be regarded as a variation of the neural networks in a deep structure. Note, the S-S structure is not considered, because SVR training is achieved by solving a large margin separator. Therefore, it is unlikely to get any advantage in concatenating a set of SVR predictions with its feature space for subsequent SVR based regression analysis.

3.3. Strength Modelling with Early and Late Fusion Strategies

As previously discussed (Sec. 2), the Strength Modelling framework can be applied in both early and late fusion strategies. Traditional early fusion combines multiple feature spaces into one single set. When integrating Strength Modelling with early fusion, the initial predictions gained from models trained on the different feature sets are also concatenated to form a new feature vector. The new feature vector is then used as the basis for the final regression analysis via a subsequent model (Fig. 2).

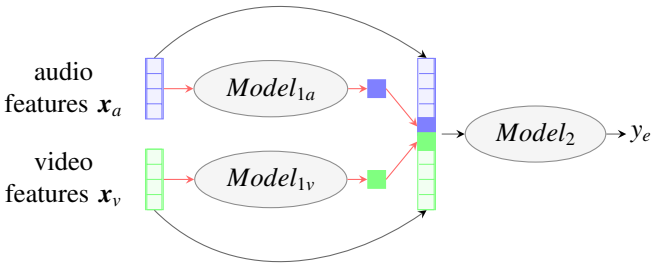


Figure 2: Strength Modelling with early fusion strategy.

Strength Modelling can also be integrated with late fusion using three different approaches, i. e., (i) modality-based, (ii) model-based, and (iii) modality- and model-based (Fig. 3). *Modality*-based fusion combines the decisions from multiple independent modalities (i. e., audio and video in our case) with the same regression model; whilst *model*-based approach fuses the decisions from multiple different models (i. e., SVR and BLSTM-RNN in our case) within the same modality; and *modality- and model*-based approach is the combination of the above two approaches, regardless of which modality or model is employed. For all three techniques the fusion weights are

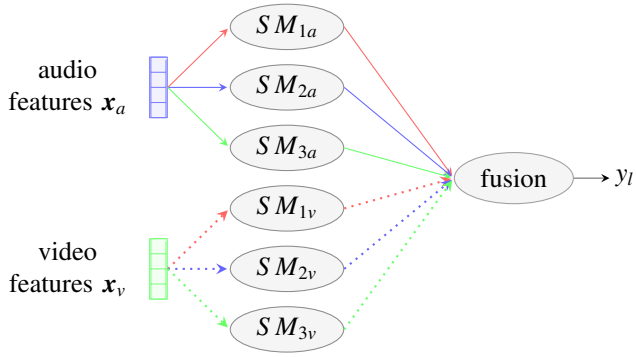


Figure 3: Strength Modelling (SM) with late fusion strategy. Fused predictions are from multiple independent modalities with the same model (denoted by the red, green, or blue lines), multiple independent models within the same modality (denoted by the solid or dotted lines), or the combination.

learnt using a linear regression model:

$$y_l = \epsilon + \sum_{i=1}^{N} \gamma_i \cdot y_i, \tag{1}$$

where y_i denotes the original prediction of the model i from N available ones; ϵ and γ_i are the bias and weights estimated on the development partition; and y_l is the final prediction.

4. Selected Databases and Features

For the transparency of experiments, we utilised the widely used multimodal continuously labelled affective databases – RECOLA [28] and SEMAINE [20], which have been adopted as standard databases for the AudioVisual Emotion Challenges (AVEC) in 2015/2016 [27, 36] and in 2012 [31], respectively. Both databases were designed to study socio-affective behaviours from multimodal data.

4.1. Databases

4.1.1. RECOLA

The RECOLA database was recorded in the context of remote collaborative work. Spontaneous interactions were collected during resolving of a collaborative task that was performed in dyads and remotely through video conference. The corpus consists of multimodal signals, i. e., audio, video, Electro-CardioGram (ECG), and Electro-Dermal Activity (EDA), which were recorded continuously and synchronously from 27 French-speaking participants. It is worth to mention that, these subjects have different mother tongues (French, Italian, and German), which provides further diversity in the encoding of affect. In order to ensure speaker-independence, the corpus was equally divided into three partitions (training, development /validation, and test), with each partition containing nine unique recording approximately balanced for gender, age, and mother tongue of the participants.

To annotate the corpus, value- and time-continuous dimensional affect ratings in terms of arousal and valence were performed by six French-speaking raters (three males and three females) for the first five minutes of all recording sequences.

The obtained labels were then resampled at a constant frame rate of 40 ms, and averaged over all raters by considering inter-evaluator agreement, to provide a 'gold standard' [28].

4.1.2. SEMAINE

The SEMAINE database was recorded in conversations between humans and artificially intelligent agents. In the recording scenario, a user was asked to talk with four emotionally stereotyped characters, which are even-tempered and sensible, happy and out-going, angry and confrontational, and sad and depressive, respectively.

For our experiments, the 24 recordings of the Solid-Sensitive Artificial Listener (Solid-SAL) part of the database were used, in which the characters were role-played. Each recording contains approximately four character conversation sessions. This Solid-SAL part was then equally split into three partitions: a training, development, and test partition, resulting in 8 recordings and 32 sessions per partition except for the training partition that contains 31 sessions. For more information on this database, the readers are referred to [31].

All sessions were annotated in continuous time and continuous value in terms of arousal and valence by two to eight raters, with the majority annotated by six raters. Different from RECOLA, the simple mean over the obtained labels was then taken to provide a single label as 'gold standard' for each dimension.

4.2. Audiovisual Feature Sets

For the acoustic features, we used the openSMILE toolkit [5] to generate 13 LLDs, i. e., 1 log energy and 12 MFCCs, with a frame window size of 25 ms at a step size of 10 ms. Rather than the official acoustic features, MFCCs were chosen as the LLDs since preliminary testing (results not given) indicated that they were more effective in association with both RECOLA [27, 36] and SEMAINE [31]. The arithmetic mean and the coefficient of variance were then computed over the sequential LLDs with a window size of 8 s at a step size of 40 ms, resulting in 26 raw features for each functional window. Note that, for SEMAINE the window step size was set to 400 ms in order to reduce the computational workload in the machine learning process. Thus, the total numbers of the extracted segments of the training, development, and test partitions were 67.5 k, 67.5 k, 67.5 k for RECOLA, and were, respectively, 24.4 k, 21.8 k, and 19.4 k for SEMAINE.

For the visual features, we retained the official features for both RECOLA and SEMAINE. As to RECOLA, 49 facial landmarks were tracked firstly, as illustrated in Fig. 4. The detected face regions included left and right eyebrows (five points respectively), the nose (nine points), the left and right eyes (six points respectively), the outer mouth (12 points), and the inner mouth (six points). Then, the landmarks were aligned with a mean shape from stable points (located on the eye corners and on the nose region).

As features for each frame, 316 features were extracted, consisting of 196 features by computing the difference between the coordinates of the aligned landmarks and those from the mean shape and between the aligned landmark locations in the previous and the current frame, 71 ones by calculating the Euclidean distances (L2-norm) and the angles (in radians) between the points in three different groups, and another 49 ones by computing the Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame. For more details on the feature extraction process the reader is referred to [27].

Again, the functionals (arithmetic mean and coefficient of variance) were computed over the sequential 316 features within a fixed length window (8 s) that shifted forward at a rate of 40 ms. As a result, 632 raw features for each functional window were included in the geometric set. Feature reduction was also conducted by applying a *Principal Component Analysis* (PCA) to reduce the dimensionality of the geometric features, retaining 95% of the variance in the original data. The final dimensionality of the reduced video feature set is 49. It should be noted that a facial activity detector was used in conjunction with the video feature extraction; video features were not extracted for the frames where no face was detected, resulting in the number of video segments somewhat less than that of audio segments.

As to SEMAINE, 5 908 frame-level features were provided as the video baseline features. In this feature set, eight features describes the position and pose of the face and eyes, and the rest are dense local appearance descriptors. For appearance descriptors, the uniform Local Binary Patterns (LBP) were used. Specifically, the registered face region was divided into 10×10 blocks, and the LBP operator was then applied to each block (59 features per block) followed by concatenating features of all blocks, resulting to another 5 900 features.

Further, to generate features on window-level, in this paper we used the method based on max-pooling. Specifically, the maximum of features were calculated with a window size of 8 s at a step size of 400 ms, to keep consistent with the audio features. We applied PCA for feature reduction on these window-level representations and generated 112 features, retaining 95% of the variance in the original data. To keep in line with RECOLA, we selected the first 49 principal components as the final video features.

5. Experiments and Results

This section empirically evaluates the proposed Strength Modelling by large-scale experiments. We first perform Strength Modelling for the continuous affect recognition in the unimodal settings (cf. Sec. 5.2), i. e., audio or video. We then incorporate it with the early (cf. Sec. 5.3) and late (cf. Sec. 5.4) fusion strategies so as to investigate its robustness in the bimodal settings.

5.1. Experimental Set-ups and Evaluation Metrics

Before the learning process, mean and variance standardisation was applied to features of all partitions. Specifically, the global means and variances were calculated from the training set, which were then applied over the development and test sets for online standardisation.

Figure 4: Illustration of the facial landmark features extraction from RECOLA database

To demonstrate the effectiveness of the strength learning, we first carried out the baseline experiments, where the SVR or BLSTM-RNNs models were individually trained on the modalities of audio, video, or the combination, respectively. Specifically, the SVR was implemented in the LIBLINEAR toolkit [6] with linear kernel, and trained with L2-regularised L2-loss dual solver. The tolerance value of ε was set to be 0.1, and complexity (C) of the SVR was optimised by the best performance of the development set among [.00001, .00002, .00005, .0001, . . . , .2, .5, 1] for each modality and task.

For the BLSTM-RNNs, two bidirectional LSTM hidden layers were chosen, with each layer consisting of the same number of memory blocks (nodes). The number was optimised as well by the development set for each modality and task among [20, 40, 60, 80, 100, 120]. During network training, gradient descent was implemented with a learning rate of 10^{-5} and a momentum of 0.9. Zero mean Gaussian noise with standard deviation 0.2 was added to the input activations in the training phase so as to improve generalisation. All weights were randomly initialised in the range from -0.1 to 0.1. Finally, the early stopping strategy was used as no improvement of the mean square error on the validation set has been observed during 20 epochs or the predefined maximum number of training epochs (150 in our case) has been executed. Furthermore, to accelerate the training process, we updated the network weights after running every mini batch of 8 sequences for computation in parallel. The training procedure was performed with our CURRENNT toolkit [38].

Herein we adapted the following naming conventions, the models trained with baseline approaches are referred to as *individual* models, whereas the ones associated with the proposed approaches are denoted as *strength* models. For the sake of a more even performance comparison the optimised parameters of individual models (i. e., SVR or BLSTM-RNN) were used in the corresponding strength models (i. e., S-B, B-S, or B-B models).

Annotation delay compensation was also performed to compensate for the temporal delay between the observable cues, as shown by the participants, and the corresponding emotion reported by the annotators [19]. Similar to [15, 36], this delay

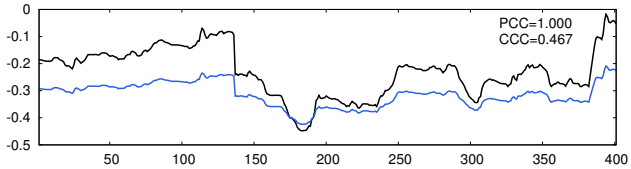


Figure 5: Comparison of PCC and CCC between two series. The black line is gold standard from RECOLA database test partition, and the blue line is generated by shifting and scaling the gold standard.

was estimated in the preliminary experiments using SVR and by maximising the performance on the development partition, while shifting the gold standard annotations back in time. As in [15, 36] we identified this delay to be four seconds which was duly compensated, by shifting the gold standard back in time with respect to the features, in all experiments presented.

Note that all fusion experiments require concurrent initial predictions from audio and visual modalities. However, as discussed in (Sec. 4.2), visual prediction cannot occur where a face has not been detected. For all fusion experiments where this occurred we replicated the initial corresponding audio prediction to fill the missing video slot.

Unless otherwise stated we report the accuracy of our systems in terms of the *Concordance Correlation Coefficient* (CCC) [27] metric:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{2}$$

where ρ is the *Pearson's Correlation Coefficient* (PCC) between two time series (e. g., prediction and gold-standard); μ_x and μ_y are the means of each time series; and σ_x^2 and σ_y^2 are the corresponding variances. In contrast to the PCC, CCC takes not only the linear correlation, but also the bias and variance between the two compared series into account. As a consequence, whereas PCC is insensitive to bias and scaling issues, CCC reflects those two variations. The value of CCC is in the range of [-1, 1], where +1 represents total concordance, −1 total discordance, and 0 no concordance at all. One may further note that, it has also been successfully used as objective function to train discriminative neural networks [39], and has been used as the official scoring metric in the last two editions of the AVEC. We further intuitively compared the difference between PCC and CCC by Fig. 5. From the figure, the obtained PCC of the two series (black and blue) is 1.000, while the obtained CCC is only 0.467 as it takes the bias of the mean and variance of the two series into account. For continuous emotion recognition, ones are often interested in not only the variation trend but also the absolute value/degree of personal emotional state. Therefore, the metric of CCC fits better for continuous emotion recognition than PCC.

In addition to CCC, results are also given in all tables in terms of *Root Mean Square Error* (RMSE), a poplar metric for regression tasks. To further access the significance level of performance improvement, a statistical evaluation was carried out over the whole predictions between the proposed and the baseline approaches by means of Fisher's r-to-z transformation [3].

Table 1: Results based on audio features only: performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) on the development and test partitions of RECOLA and SEMAINE databases from the *audio* signals. The best achieved CCC is highlighted. The symbol of $*$ indicates the significance of the performance improvement over the related individual systems.

Audio based method	RECOLA AROUSAL RMSE	RECOLA AROUSAL CCC	RECOLA VALENCE RMSE	RECOLA VALENCE CCC	SEMAINE AROUSAL RMSE	SEMAINE AROUSAL CCC	SEMAINE VALENCE RMSE	SEMAINE VALENCE CCC
a. on the *development* set								
S	.126	.714	.149	.331	.218	.399	.262	.172
B	.142	.692	.117	.286	.209	.387	.261	.117
B-S	.127	.713	.144	.348*	.206	.417*	.255	**.179**
S-B	.122	.753*	.113	.413*	.210	**.434***	.262	.172
B-B	.122	**.755***	.112	**.476***	.206	.417*	.255	.178*
b. on the *test* set								
S	.133	.605	.165	.248	.216	.397	.263	**.017**
B	.155	.625	.119	.282	.202	.317	.256	.008
B-S	.133	.606	.160	.264	.205	.332	.258	.006
S-B	.133	.665*	.117	.319*	.203	**.423***	.262	**.017**
B-B	.133	**.666***	.123	**.364***	.205	.332*	.258	.006

Table 2: Results based on visual features only: performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) on the development and test partitions of RECOLA and SEMAINE databases from the *video* signals. The best achieved CCC is highlighted. The symbol of $*$ indicates the significance of the performance improvement over the related individual systems.

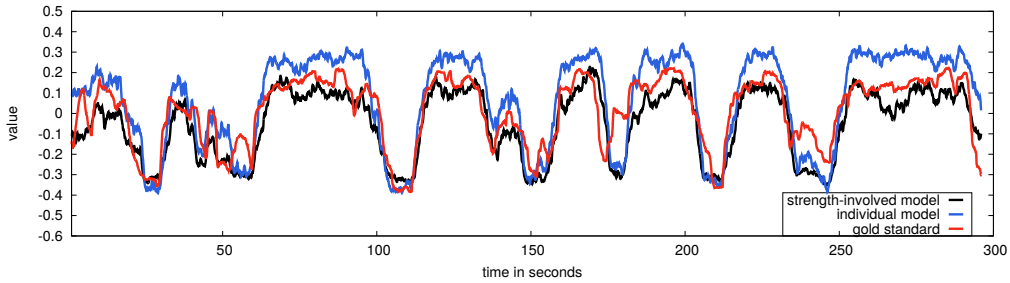
Video based method	RECOLA AROUSAL RMSE	RECOLA AROUSAL CCC	RECOLA VALENCE RMSE	RECOLA VALENCE CCC	SEMAINE AROUSAL RMSE	SEMAINE AROUSAL CCC	SEMAINE VALENCE RMSE	SEMAINE VALENCE CCC
a. on the *development* set								
S	.197	.120	.139	.456	.249	.241	.253	.393
B	.184	.287	.110	.478	.224	.232	.247	.332
B-S	.183	.292	.110	**.592***	.222	.250	.252	.354
S-B	.186	**.350***	.118	.510*	.231	**.291***	.242	**.405**
B-B	.185	.344*	.113	.501*	.222	.249*	.256	.301
b. on the *test* set								
S	.186	.193	.156	.381	.279	.112	.278	.115
B	.183	.193	.122	.394	.240	.112	.275	.063
B-S	.176	**.265***	.130	.464*	.235	.072	.285	.043
S-B	.186	.196	.121	**.477***	.249	**.125**	.284	.068
B-B	.197	.184	.120	.459*	.235	.072	.255	**.158***

Unless stated otherwise, a p value less than .05 indicates significance.

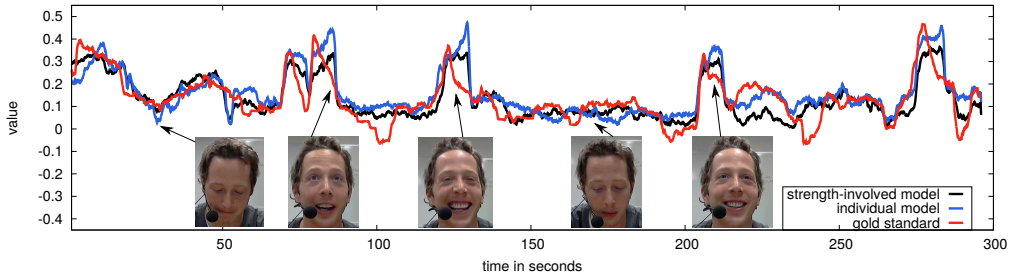
5.2. *Affect Recognition with Strength Modelling*

Table 1 displays the results (RMSE and CCC) obtained from the strength models and the individual models of SVR and BLSTM-RNN on the development and test partitions of RECOLA and SEMAINE databases from the *audio*. As can be seen, the three Strength Modelling set-ups either matched or outperformed their corresponding individual models in most cases. This observation implies that the advantages of each model (i. e., SVR and BLSTM-RNN) are enhanced via Strength Modelling. In particular the performance of the BLSTM model, for both arousal and valence, was significantly boosted by the inclusion of SVR predictions (S-B) on the development and test sets. We speculate this improvement could be due to the initial SVR predictions helping the subsequent RNN avoid local minima.

Similarly, the B-S combination brought additional performance improvement for the SVR model (except the valence case of SEMAINE), although not as obvious as for the S-B model. Again, we speculate that the temporal information leveraged by the BLSTM-RNN is being exploited by the successive SVR model. The best results for both arousal and valence dimensions were achieved with the framework of B-B for RECOLA, which achieved relative gains of 6.5 % and 29.1 % for arousal and valence respectively on the test set when compared to the single BLSTM-RNN model (B). This indicates



(a) *arousal* prediction via *audio* signals



(b) *valence* prediction via *video* signals

Figure 6: Automatic prediction of arousal via audio signals (a) and valence via video signals (b) obtained with the best settings of the *strength*-involved models and *individual* models for a subject from the test partition on RECOLA database.

there are potential benefits for audio based affect recognition by the deep structure formed by combining two BLSTM-RNNs using the Strength Modelling framework. Additionally, one can observe that there is no much performance improvement by applying Strength Modelling in the case of the valence recognition of SEMAINE. This might be attribute to the poor performance of the baseline systems, which can be regarded as noise and possibly not able to provide useful information for the other models.

The same set of experiments were also conducted on the video feature set (Table 2). As for valence, the highest CCC obtained on test set achieves at .477 using the S-B model for RECOLA and at .158 using the B-B model for SEMAINE. As expected, we observe that the models (individual or strength) trained using only acoustic features is more efficient for interpreting the dimension of arousal rather than valence. Whereas, the opposite observation is seen for models trained only on the visual features. This finding is in agreement with similar results in the literature [8, 10, 26].

Additionally, Strength Modelling achieved comparable or superior performance to other state-of-the-art methods applied on the RECOLA database. The OA-RVM model was used in [15, 18], and the reported performance in terms of CCC, with audio features on the development set, was .689 for arousal [15], and .510 for valence using video features [18]. We achieved .755 with audio features for arousal, and .592 with video features for valence with the proposed Strength Modelling framework, showing the interest of our method.

To further highlight advantages of Strength Modelling, Fig. 6 illustrates the automatic predictions of arousal via audio signals (a) and valence via video signals (b) obtained with the best settings of the strength models and the individual models frame by frame for a single test subject from RECOLA. Note that, similar plots were observed for the other subjects in the test set. In general, the predictions generated by the proposed Strength Modelling approach are closer to the gold standard, which consequently contributes to better results in terms of CCC.

5.3. Strength Modelling Integrated with Early Fusion

Table 3 shows the performance of both the individual and strength models integrated with the early fusion strategy. In most cases, the performance of the *individual* models of either SVR or BLSTM-RNN was significantly improved with the fused feature vector for both arousal and valence dimensions in comparison to the performance with the corresponding individual models trained only on the unimodal feature sets (Sec. 5.2) in most cases for both RECOLA and SEMAINE datasets.

For the strength model systems, the early fusion B-S model generally outperformed the equivalent SVR model, and the structure of S-B outperformed the equivalent BLSTM model. However, the gain obtained by Strength Modelling with the early fused features is not as obvious as that with individual models. This might be due to the higher dimensions of the fused feature sets which possibly reduce the weight of the predicted features.

5.4. Strength Modelling Integrated with Late Fusion

This section aims to explore the feasibility of integrating Strength Modelling into three different late fusion strategies: modality-based, model-based, and the combination (see Sec. 3.3). A comparison of the performance of different fusion approaches, with or without Strength Modelling, is presented in Table 4. For the systems *without* Strength Modelling for RECOLA, one can observe that best individual model test

Table 3: Early fusion results on RECOLA and SEMAINE databases: performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) with *early fusion strategy* on the development and test partitions of RECOLA and SEMAINE databases. The best achieved CCC is highlighted. The symbol of $*$ indicates the significance of the performance improvement over the related individual systems.

Early Fusion	RECOLA				SEMAINE			
	AROUSAL		VALENCE		AROUSAL		VALENCE	
method	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC
a. on the *development* set								
S	.121	**.728**	.113	.544	.213	.392	.252	**.436**
B	.132	.700	.109	.513	.217	.354	.257	.205
B-S	.122	.727	.118	**.549**	.210	.374	.239	.363
S-B	.127	.712	.096	.526	.208	**.423***	.253	.397
B-B	.126	.718*	.095	.542*	.210	.421*	.241	.361*
b. on the *test* set								
S	.132	.610	.139	.463	.224	.304	.292	.057
B	.148	.562	.114	.476	.204	.288	.244	.127
B-S	.132	.610	.121	**.520***	.204	.328*	.264	.063
S-B	.144	.616*	.112	.473	.198	**.408***	.275	**.144***
B-B	.143	**.618***	.114	.499*	.220	.307*	.265	.060

set performances, .625 and .394, for arousal and valence respectively (Sec. 5.2) were boosted to .671 and .405 with the modality-based late fusion approach, and to .651 and .497 with the model-based late fusion approach. These results were further promoted to .664 and .549 when combining the modality- and model-based late fusion approaches. This result is in line with other results in the literature [27, 15], and again confirms the importance of multimodal fusion for affect recognition. However, similar observation can only been seen on the validation set for SEMAINE, which might be due to the huge mismatch between the validation and test partitions.

Interestingly when incorporating Strength Modelling into late fusion we can observe significant improvements over the corresponding non-strength set-ups. This finding confirms the effectiveness and the robustness of the proposed method for multimodal continuous affect recognition. In particular, the best test results of RECOLA, .685 and .554, were obtained by the strength models integrated with the modality- and model-based late fusion approach. This arousal result matches the performance with the AVEC 2016 affect recognition subchallenge baseline system, .682, which was obtained using a late fusion strategy involving eight feature sets [36].

As for SEMAINE, although obvious performance improvement can be seen on the development set, a similar observation can not be observed on the test set. This finding is possibly attributed to the mismatch between the development set and the test set, since all parameters of the training models were optimised on the development set. However, these parameters are not fit for the test set anymore.

Further, for a comparison with the OA-RVM system, we applied the same fusion system as used in [15], with only audio and video features. The results are shown in Table 4 and 5 for the RECOLA and SEMAINE database, respectively. It can be seen that, for both databases, the proposed methods outperform the OA-RVM technique, which further confirms the efficiency of the proposed Strength Modelling method.

In general, to provide an overview of the contributions of Strength Modelling to the continuous emotion recognition, we averaged the relative performance improvement of Strength Modelling over RECOLA and SEMAINE for arousal and valence recognition. The corresponding results from four cases (i. e., audio only, video only, early fusion, and late fusion) are displayed in Fig. 7. From the figure, one can observe an obvious performance improvement gained by Strength Modelling, except for the late fusion framework. This particular case is highly attributed to the mismatch between validation and test sets of SEMAINE as aforementioned, as all parameters of the training models were optimised on the development set. Employing some state-of-the-art generation techniques like dropout for training neural networks might help to tackle this problem in the future.

6. Conclusion and Future Work

This paper proposed and investigated a novel framework, *Strength Modelling*, for continuous audiovisual affect recognition. Strength Modelling concatenates the strength of an initial model, as represented by its predictions, with the original features to form a new feature set which is then used as the basis for regression analysis in a subsequent model.

To demonstrate the suitability of the framework, we jointly explored the benefits from two state-of-the-art regression models, i. e., Support Vector Regression (SVR) and Bidirectional Long Short-Term Memory Recurrent Neural Network (BLSTM-RNN), in three different Strength Modelling structures (SVR-BLSTM, BLSTM-SVR, BLSTM-BLSTM). Further, these three structures were evaluated in both unimodal settings, using either audio or video signals, and the bimodal settings where early fusion and late fusion strategies were integrated.

Table 4: Late fusion results on the RECOLA database: performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) with *late fusion strategies* (i. e., modality-based, model-based, or the combination) on the development and test partitions of RECOLA database. The best achieved CCC is highlighted. The symbol of $*$ indicates the significance of the performance improvement over the related individual systems.

Late Fusion	RECOLA							
	AROUSAL				VALENCE			
	Dev.		Test		Dev.		Test	
fusion type	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC
a. *modality*-based								
A+V; S	.117	.777	.134	.654	.128	.493	.149	.386
A+V; B	.126	.736	.134	.671	.104	.475	.113	.405
A+V; B-S	.114	**.791***	.130	.668	.090	**.664***	.105	**.542***
A+V; S-B	.117	.778	.128	**.681***	.096	.586*	.105	.495*
A+V; B-B	.117	.779*	.130	.680*	.095	.601*	.106	.506*
b. *model*-based								
A; S+B	.119	.771	.132	.651	.112	.335	.117	.284
V; S+B	.179	.230	.172	.184	.096	.588	.110	**.497**
A; (B-S)+(S-B)+(B-B)	.117	**.778***	.132	**.664***	.108	**.409***	.120	**.303***
V; (B-S)+(S-B)+(B-B)	.171	**.344***	.171	**.222***	.095	**.599***	.111	.477
c. *modality*- and *model*-based								
A+V; S+B	.113	.795	.130	.664	.089	.670	.107	.549
A+V; (B-S)+(S-B)+(B-B)	.110	**.808***	.127	**.685***	.088	**.671**	.103	**.554**
state-of-the-art method								
OA-RVM	.135	.725	.150	.612	.171	.384	.169	.392

Table 5: Late fusion results on the SEMAINE database: performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) with *late fusion strategies* (i. e., modality-based, model-based, or the combination) on the development and test partitions of SEMAINE database. The best achieved CCC is highlighted. The symbol of $*$ indicates the significance of the performance improvement over the related individual systems.

Late Fusion	SEMAINE							
	AROUSAL				VALENCE			
	Dev.		Test		Dev.		Test	
fusion type	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC
a. *modality*-based								
A+V; S	.205	.416	.205	**.370**	.231	.422	.271	.097
A+V; B	.202	.439	.210	.313	.240	.351	.276	.055
A+V; B-S	.200	**.460***	.211	.305	.238	.369	.271	.033
A+V; S-B	.201	.445	.207	.368	.231	**.424**	.278	.062
A+V; B-B	.200	**.460***	.211	.304	.242	.336	.257	**.099***
b. *model*-based								
A; S+B	.207	.394	.201	.348	.254	.212	.261	**.021**
V; S+B	.222	.238	.229	**.125**	.237	.376	.273	**.096**
A; (B-S)+(S-B)+(B-B)	.204	**.420***	.202	**.364***	.253	**.226**	.262	.014
V; (B-S)+(S-B)+(B-B)	.221	**.246**	.231	.084	.235	**.390***	.300	.036
c. *modality*- and *model*-based								
A+V; S+B	.201	.447	.206	**.353**	.235	.395	.277	**.054**
A+V; (B-S)+(S-B)+(B-B)	.198	**.470***	.207	.346	.224	**.477***	.301	.026
state-of-the-art method								
OA-RVM	.253	.433	.247	.346	.312	.315	.351	.021

Results gained on the widely used RECOLA and SEMAINE databases indicate that Strength Modelling can match or outperform the corresponding conventional individual models when performing affect recognition. An interesting observation was that, among our three different Strength Modelling set-ups no one case significantly outperformed the others. This demonstrates the flexibility of the proposed framework, in terms of being able to work in conjunction with different combination of

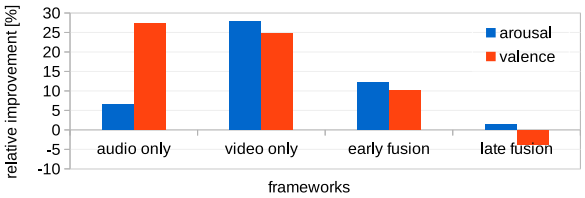


Figure 7: Averaged relative performance improvement (in terms of CCC) cross RECOLA and SEMAINE for arousal and valence recognition. The performance of the Strength Modelling was compared with the best individual systems in the case of audio only, video only, early fusion, and late fusion frameworks.

regression strategies.

A further advantage of Strength Modelling is that, it can be implemented as a plug-in for use in both early and late fusion stages. Results gained from an exhaustive set of fusion experiments confirmed this advantage. The best Strength Modelling test set results on the RECOLA dataset, .685 and .554, for arousal and valence respectively were obtained using Strength Modelling integrated into a modality- and model-based late fusion approach. These results are much higher than the ones obtained from other state-of-the-art systems. Moreover, on the SEMAINE dataset, competitive results can also be obtained.

There is a wide range of possible future research direction associated with Strength Modelling to build on this initial set of promising results. First, only two widely used regression model were investigated in the present article for affect recognition. Much of our future efforts will concentrate around assessing the suitability of more other regression approaches (e. g., Partial Least Squares Regression) for use in the framework. Investigating a more general rule of what kind of models can be implemented together in the framework help to expand its application. In addition, it is interesting to extend the framework widely and deeply. Second, motivated by the work in [17], we will also combine the original features with the predictions from different modalities (integrating the predictions based on audio features with the original video features for a final arousal or valence prediction), rather than from different models only. Furthermore, we also plan to generalise the promising advantages offered by Strength Modelling, by evaluating its performance on other behavioural regression tasks.

Acknowledgements

This work was supported by the EU's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA) and the EC's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu). We further thank the NVIDIA Corporation for their support of this research by Tesla K40-type GPU donation.

References

[1] Chang, C.-Y., Chang, C.-W., Zheng, J.-Y., Chung, P.-C., Dec 2013. Physiological emotion analysis using support vector regression. Neurocomputing 122, 79–87.

[2] Chao, L., Tao, J., Yang, M., Li, Y., Wen, Z., 2015. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proc. the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC). Brisbane, Australia, pp. 65–72.

[3] Cohen, J., Cohen, P., West, S. G., Aiken, L. S., 2013. Applied multiple regression/correlation analysis for the behavioral sciences. Routledge, Abingdon, UK.

[4] Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., Vapnik, V., 1997. Support vector regression machines. Denver, CO, pp. 155–161.

[5] Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE – the Munich versatile and fast open-source audio feature extractor. In: Proc. ACM International Conference on Multimedia (ACM MM). Florence, Italy, pp. 1459–1462.

[6] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., Jun 2008. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874.

[7] Graves, A., Schmidhuber, J., Jul 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18 (5), 602–610.

[8] Gunes, H., Pantic, M., Jan 2010. Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions 1 (1), 68–99.

[9] Gunes, H., Piccardi, M., Feb 2009. Automatic temporal segment detection and affect recognition from face and body display. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (1), 64–84.

[10] Gunes, H., Schuller, B., Feb 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing 31 (2), 120–136.

[11] Gunn, S. R., May 1998. Support vector machines for classification and regression. Tech. Rep. 14, School of Electronics and Computer Science, University of Southampton, Southampton, England.

[12] He, L., Jiang, D., Yang, L., Pei, E., Wu, P., Sahli, H., 2015. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proc. the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC). Brisbane, Australia, pp. 73–80.

[13] Hermansky, H., Ellis, D. P. W., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMM systems. In: Proc. IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP). Istanbul, Turkey, pp. 1635–1638.

[14] Hochreiter, S., Schmidhuber, J., Nov 1997. Long short-term memory. Neural Computation 9 (8), 1735–1780.

[15] Huang, Z., Dang, T., Cummins, N., Stasak, B., Le, P., Sethu, V., Epps, J., 2015. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In: Proc. the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC). Brisbane, Australia, pp. 41–48.

[16] Kumar, N., Gupta, R., Guha, T., Vaz, C., Van Segbroeck, M., Kim, J., Narayanan, S. S., 2014. Affective feature design and predicting continuous affective dimensions from music. In: Proc. MediaEval. Barcelona, Spain.

[17] Kursun, O., Seker, H., Grgen, F., Aydin, N., Favorov, O. V., Sakar, C. O., 2009. Parallel interacting multiview learning: An application to prediction of protein sub-nuclear location. In: Proc. 9th International Conference on Information Technology and Applications in Biomedicine (ITAB). Larnaca, Cyprus, pp. 1–4.

[18] Manandhar, A., Morton, K. D., Torrione, P. A., Collins, L. M., Jan 2016. Multivariate output-associative RVM for multi-dimensional affect predictions. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering 10 (3), 408–415.

[19] Mariooryad, S., Busso, C., April 2015. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. IEEE Transactions on Affective Computing 6 (2), 97–108.

[20] McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M., Jan 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on Affective Computing 3 (1), 5–17.

[21] Nicolaou, M. A., Gunes, H., Pantic, M., Apr 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing 2 (2), 92–105.

[22] Nicolaou, M. A., Gunes, H., Pantic, M., Mar 2012. Output-associative rvm regression for dimensional and continuous emotion prediction. Image and Vision Computing 30 (3), 186–196.

[23] Pantic, M., Rothkrantz, L. J. M., Sep 2003. Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE 91 (9), 1370–1390.

[24] Petridis, S., Pantic, M., Jan 2016. Prediction-based audiovisual fusion for classification of non-linguistic vocalisations. IEEE Transactions on Affective Computing 7 (1), 45–58.

[25] Qiu, X., Zhang, L., Ren, Y., Suganthan, P. N., Amaratunga, G., 2014. Ensemble deep learning for regression and time series forecasting. In: Proc. Computational Intelligence in Ensemble Learning (CIEL). Orlando, FL, pp. 1–6.

[26] Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D., Schuller, B., Nov 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters 66, 22–30.

[27] Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M., 2015. AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In: Proc. the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC). Brisbane, Australia, pp. 3–8.

[28] Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D., 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proc. EmoSPACE (FG). Shanghai, China, pp. 1–8.

[29] Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., Rigoll, G., 2005. Speaker independent speech emotion recognition by ensemble classification. In: Proc. IEEE International Conference on Multimedia and Expo (ICME). Amsterdam, Netherland, pp. 864–867.

[30] Schuller, B., Rigoll, G., Lang, M., 2003. Hidden markov model-based speech emotion recognition. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. I. Hong Kong, China, pp. I401–I404.

[31] Schuller, B., Valster, M., Eyben, F., Cowie, R., Pantic, M., 2012. AVEC 2012: the continuous audio/visual emotion challenge. In: Proc. the 14th ACM International Conference on Multimodal Interaction (ICMI). Nara, Japan, pp. 449–456.

[32] Soleymani, M., Aljanaki, A., Yang, Y.-H., Caro, M. N., Eyben, F., Markov, K., Schuller, B. W., Veltkamp, R., Weninger, F., Wiering, F., 2014. Emotional analysis of music: A comparison of methods. In: Proc. ACM International Conference on Multimedia. New York, NY, pp. 1161–1164.

[33] Soleymani, M., Asghari-Esfeden, S., Fu, Y., Pantic, M., Jan 2016. Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Transactions on Affective Computing 7 (1), 17–28.

[34] Soleymani, M., Asghari-Esfeden, S., Pantic, M., Fu, Y., 2014. Continuous emotion detection using EEG signals and facial expressions. In: Proc. IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6.

[35] Tian, L., Moore, J. D., Lai, C., 2015. Emotion recognition in spontaneous and acted dialogues. In: Proc. Affective Computing and Intelligent Interaction (ACII). Xi'an, China, pp. 698–704.

[36] Valstar, M. F., Gratch, J., Schuller, B. W., Ringeval, F., Lalanne, D., Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M., 2016. AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. In: Proc. the 6th International Workshop on Audio/Visual Emotion Challenge. Amsterdam, The Netherlands, pp. 3–10.

[37] Wei, J., Pei, E., Jiang, D., Sahli, H., Xie, L., Fu, Z., 2014. Multimodal continuous affect recognition based on LSTM and multiple kernel learning. In: Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Siem Reap, Cambodia, pp. 1–4.

[38] Weninger, F., Bergmann, J., Schuller, B., Jan 2015. Introducing CURRENNT: The munich open-source cuda recurrent neural network toolkit. Journal of Machine Learning Research 16 (1), 547–551.

[39] Weninger, F., Ringeval, F., Marchi, E., Schuller, B., 2016. Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In: Proc. International Joint Conference on Artificial Intelligence (IJCAI). New York, NY, pp. 2196–2202.

[40] Wöllmer, M., Eyben, F., Graves, A., Schuller, B., Rigoll, G., Apr 2010. Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. Cognitive Computation 2 (3), 180–190.

[41] Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R., 2008. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. INTERSPEECH. Brisbane, Australia, pp. 597–600.

[42] Wöllmer, M., Eyben, F., Schuller, B., Sun, Y., Moosmayr, T., Nguyen-Thien, N., 2009. Robust in-car spelling recognition-a tandem BLSTM-HMM approach. In: Proc. INTERSPEECH. Brighton, UK, pp. 2507–2510.

[43] Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., Rigoll, G., Feb 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing 31 (2), 153–163.

[44] Wolpert, D. H., Dec 1992. Stacked generalization. Neural networks 5 (2), 241–259.

[45] Wu, C.-H., Lin, J.-C., Wei, W.-L., 2014. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA Transactions on Signal and Information Processing 3, e12, 18 pages.

[46] Yang, Y. H., Lin, Y. C., Su, Y. F., Chen, H. H., Feb 2008. A regression approach to music emotion recognition. IEEE Transactions on Audio, Speech, and Language Processing 16 (2), 448–457.

[47] Zeng, Z., Pantic, M., Roisman, G. I., Huang, T. S., Jan 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (1), 39–58.

[48] Zhang, Z., Pinto, J., Plahl, C., Schuller, B., Willett, D., Aug 2014. Channel mapping using bidirectional long short-term memory for dereverberation in hand-free voice controlled devices. IEEE Transactions on Consumer Electronics 60 (3), 525–533.

[49] Zhang, Z., Ringeval, F., Dong, B., Coutinho, E., Marchi, E., Schuller, B., 2016. Enhanced semi-supervised learning for multimodal emotion recognition. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China, pp. 5185–5189.

[50] Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., Schuller, B., 2016. Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks. In: Proc. INTERSPEECH. San Francisco, CA, pp. 3593–3597.