Ensemble Methods for Continuous Affect Recognition: Multi-modality, Temporality, and Challenges

Markus Kächele
Institute of Neural Information Processing
Ulm University, Germany
markus.kaechele@uni-ulm.de

Patrick Thiam
Institute of Neural Information Processing
Ulm University, Germany
patrick.thiam@uni-ulm.de

Günther Palm
Institute of Neural Information Processing
Ulm University, Germany
guenther.palm@uni-ulm.de

Friedhelm Schwenker
Institute of Neural Information Processing
Ulm University, Germany
friedhelm.schwenker@uni-ulm.de

Martin Schels
Institute of Neural Information Processing
Ulm University, Germany
martin.schels@uni-ulm.de

ABSTRACT

In this paper we present a multi-modal system based on audio, video and bio-physiological features for continuous recognition of human affect in unconstrained scenarios. We leverage the robustness of ensemble classifiers as base learners and refine the predictions using stochastic gradient descent based optimization on the desired loss function. Furthermore we provide a discussion about pre- and post-processing steps that help to improve the robustness of the regression and subsequently the prediction quality.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords

AVEC 2015; affect recognition; multi-modal fusion

1. INTRODUCTION

The automatic recognition of human emotions is basically solved for constrained laboratory conditions with acted affective databases [23, 15]. However for real world conditions with visual occlusion and affective states of low expressiveness, the recognition performances commonly drop significantly [22, 12, 21].

The AVEC challenges aim to create a benchmarks to evaluate classification systems that are capable of robust affect recognition beyond laboratory conditions. The data set of the 2015 edition provides 4 different modalities for classification. While it may be challenging or technically infeasible to create one single classifier, usually multiple classifier systems or related multi-modal classification architectures are created to deal with different input channels such as audio, video or even bio-physiology. Predictions that are based on multiple modalities that support each other are generally more robust than those that are based on individual modalities alone.

In order to create a valuable contribution to AVEC 2015, we briefly review the submissions of last year's challenge. The baseline was set using ϵ-SVR [25], reaching an average correlation coefficient on the test set of 0.419. A great variety of learning algorithms have been used by the challenge participants. Kaya et al. used extreme learning machines and obtained a mean correlation coefficient of 0.393 [16]. Further, Gupta et al. used SVR based on additional video and text based features and thus rendered an average correlation coefficient of 0.488 [10]. Chao et al. used deep learning techniques and a temporal pooling mechanism to obtain a mean correlation coefficient of 0.550 [4]. The best entry for the affect sub-challenge of the 2014 edition of AVEC was submitted by Kächele et al. with an average correlation coefficient of 0.595 [13]. In that work, novel task dependent proto-labels were developed that were constructed using ϵ-SVR and Eigenvalue decomposition. Additionally a personalization step by clustering user groups based on global features was included to improve the predictions.

The remainder of this work is organized as follows. In the next Section, we discuss the current iteration of the challenge and how it stands out from previous iterations. In Section 3, the dataset and feature extraction pipeline are explained. We propose a feedforward network that directly optimizes concordance correlation in Section 4. Extensive experiments are presented in Section 5 together with a discussion of our results and findings. We conclude the paper in Section 7.

AVEC'15, October 26, 2015, Brisbane, Australia.

ACM 978-1-4503-3743-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2808196.2811637.

2. THE SAME PROCEDURE AS EVERY YEAR?

The AVEC challenge has been around since 2011 and evolved over the years to become one of the most well known benchmarks in affective computing. To maintain steady interest in the challenge, it is slightly altered every year. Objectively the changes that have been implemented for this year's iteration are some of the biggest the challenge has seen so far since the introduction of fully continuous label traces.

One important adaptation is using Lin's concordance correlation coefficient (CCC) [17] rather than Pearson's correlation coefficient to measure performance. Thus, predictions that are shifted in value are punished in contrast to what it was before. A further rather small but nevertheless interesting point is that for the evaluation of the test data the individual sequences are concatenated and based on this the CCC is calculated. This is a different policy than in previous editions where the average is taken over the individual performances.

Another important change is an apparent increase of the amplitudes of the label traces during one session as seen in Figure 1. The figure shows the mean absolute values of the provided labels together with their variances over the sequences of the training and validation sets of the 2012, 2014 and 2015 editions of the challenge, respectively. It is clearly observable that most of the label traces in the previous editions are mostly dominated by a transient phase that is needed for the labeler to move away from the initial starting point. Even though the continuous labels still start from a uniform starting point the whole curves are by far less influenced by it. The origin of this is that there seems to be a higher expressiveness in the recordings which renders larger amplitudes for the continuous annotation traces.

Further, the social group of which the test subjects were recruited from differs significantly at least compared to the last two challenges where the recordings were made in a clinical context. The main target of the last challenges' data sets was not so much the recognition of affective states but rather the investigation of severity of depression. A further difference to previous challenges which might be the main reason for the increased expressiveness is the fact that the data is recorded in a human-human interaction scenario. In contrast to that the subjects were interacting with an artificial avatar or reading from a computer screen in previous editions. This contradicts to some degree the expectation to study and improve human-computer interaction as it is advertised in the guidelines of the challenge.

3. DATASET AND FEATURES

The data collection that is used in the AVEC2015 competition is the RECOLA database that was recorded at the University of Fribourg, Switzerland [20]. It comprises 27 sessions of length 5 minutes each, which consist of 4 different channels: audio, video, electrocardiogram and electrodermal activity. The two affective dimensions "arousal" and "valence" were manually annotated using a slider-based label tool. Each recording was annotated by 6 native French speakers. The average of these individual ratings is used as true label here. A sample screen-shot of the data is shown in Figure 2.

Figure 2: Exemplary recording situation for the AVEC 2015 data set.

The data set is provided together with a set of pre-calculated features which we incorporated into our classifiers (see [19] for details). Apart from that, a number of additional features were extracted from the data that are briefly described in the following.

3.1 Audio Features

The extraction of **linear predictive coding coefficients (LPC)** is an auto regressive approach, where the n^{th} sample of a time series is approximated using a function of the p preceding samples [1]: LPC are still widely applied in speech processing, for example in speech recognition and speech synthesis. One reason for this is that they are easily computed as no Fourier analysis has to be conducted. For the speech classification in this work, 8 LPC were computed for time windows of 32 ms length with an offset of 16 ms.

The **mel frequency cepstral coefficients (MFCC)** [7] are computed by first obtaining the short term power spectrum using the discrete Fourier transformation from windows of speech. Then a conversion of the spectrum to the mel scale is carried out. After triangular bandpass filtering, the logarithm of the powers is computed. The final coefficients are obtained by computing the discrete cosine transform of the resulting signal.

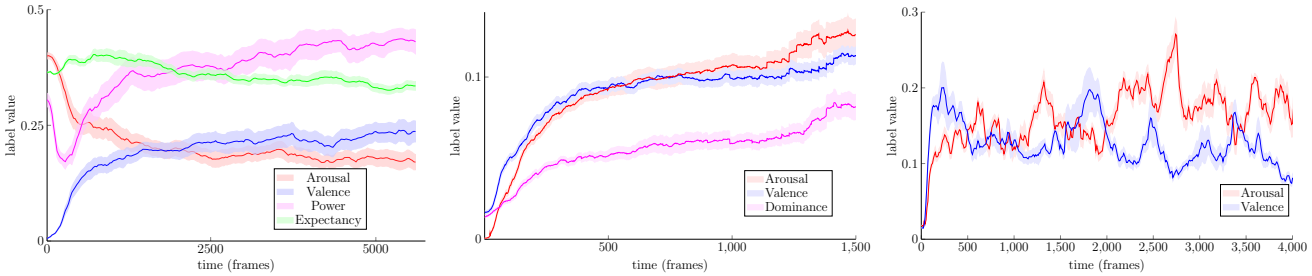
Using the triangular filters in mel-scale, i.e., higher densities of filters with smaller bandwidths for lower frequencies than for higher frequencies, is an idea that is inspired by the human ear [7].

For the experiments here 20 MFCC coefficients per time window of length 32 ms with an offset of 16 ms are extracted from the audio signal.

Log frequency power coefficients [18] are computed using a log filter bank ranging from 200 Hz to 4 kHz. As it is the case for MFCC, the filter bank is designed such that it follows the human frequency resolution. For each filter, the signals are preprocessed using a Hamming window, followed by DFT to obtain the spectrum $S(m)$ for frequency band m. The final LFPC coefficients are computed by $LFPC(m) = \frac{10\log_{10}(S(m))}{N_m}$ where N_m is the number of spectral components in the respective filter.

3.2 Video Features

Local binary patterns in three orthogonal planes (LBP-TOP) [9] are an efficient approximation of space-time or volumetric LBP. Those variants of local binary pat-



(a) Mean absolute labels for AVEC2012 (b) Mean absolute labels for AVEC2014 (c) Mean absolute labels for AVEC2015

Figure 1: Mean absolute labels for three different editions of the AVEC challenge. It can be seen that the expressivity of the labels are increased in AVEC2015 in relation to the transient phase compared to the 2012 and 2014 editions of the event.

terns are attractive because additionally to textural information, they encode information about the dynamics of a sequence. LBP-TOP is computed using the standard local binary patterns algorithm for planes that are oriented in the $x-y$, $x-t$ and $y-t$ of the space-time volume. In this work, the LBP-TOP features were used that have been provided by the challenge organizers.

We further use another set of LBP-TOP coefficients computed on a window of one second, with the facial region divided into 2×2 blocks that are overlapping by 25%.

Histograms of oriented gradients (HOG) [6] are used to capture structural information of small neighborhoods in an image. They are computed by processing the image with a Canny edge detector, computing the gradients using horizontal and vertical Sobel operators and binning the resulting gradient direction into one of K bins equally spaced across the interval $[0, 2\pi)$.

In this work the aligned facial image is partitioned into 32×32 pixel windows and the number of bins in the histograms was set to 9. This renders a feature vector of dimension 324.

Pyramids of histograms of oriented gradients in three orthogonal planes (PHOG-TOP) are an extension of the original HOG. PHOG [2] combines spatial information with the distribution of image gradient orientations by introducing a multi-resolution scheme using an image pyramid. On every pyramid level l each dimension is divided into 2^l cells. Then, for every cell, a HOG descriptor is computed. The final PHOG descriptor is a concatenation of all HOG descriptors over every pyramid level. By treating PHOG like LBP above, a new descriptor can be created by computing the result for the defined planes in the space time volume. Consequently, the descriptor is from now on referred to as PHOG-TOP.

3.3 Biophysiology

The feature extraction from the bio-physiological channels was preceded by bandpass filtering, detrending and artifact correction.

Electrocardiogram: The processing of the ECG channel included piecewise linear detrending followed by detection of the so called QRS complexes (characteristic waveform of the heart beats). This was accomplished by detection and pooling of local maxima over short time windows. The features included the *amplitudes* of P, Q, R, S and T points, the *time delay* between points and the *angles* of the Q and S valleys. These features are denoted as **ECG-PQRST** in the following.

The next group of features is based on the wavelet decomposition of the signal. Necessary aligning was done based on the detected R peaks. The window length used for the segmentation was 650 ms, including 250 ms before and 400 ms after the R peak. After practical experiments a Daubechies wavelet of order 8 was chosen. Four level wavelet decomposition was applied to obtain the desired features. The approximation coefficients of the wavelet decomposition a_1 were considered as the features of each heart beat since the detail coefficients contain mostly noise and the high frequency components of the signal. Ultimately the *mean* of the approximation coefficients of each heart beat within the entire window were considered as features for the recognition task. These features are denoted as **ECG-Wavelet**.

Skin conductance level: The skin conductance level indicates the activity of the sweat glands in the skin and is directly controlled by the sympathetic nervous system. For the SCL channel, a set of features has been computed that was designed for feature extraction from electromyography (EMG). For more information about the feature extraction process, the reader is referred to [14].

Additionally, statistical features such as the *skewness*, *kurtosis*, *ratio* between maximum and minimum, *number* and *mean amplitudes* of *SCR occurrences*, *temporal slope* of the signal, and *normalized length density* were computed.

Subsequently, the following features have been extracted: *mean* and *standard deviation* of the signal and the first and second derivatives. Based on Welch's power spectrum density estimation the *ratio* of low frequency to very low frequency was computed.

The amplitude of the signal gave rise to the features *peak amplitude* (value of highest peak) and *range* (difference between highest and lowest value), as well as the *root of the squared mean* and the *mean of the absolute* signal. These features are denoted as **SCL** in the following.

The *central, mode and mean frequencies* are measures of the rate of vibrations of facial muscles, while the *bandwidth* represents its variability. Here it is applied it to extract information from the characteristic SCL trajectory.

Approximation entropy, *sample entropy*, *fuzzy entropy* [5], and *Shannon entropy* are introduced to measure the irregularity and unpredictability of the signals. The mentioned features are referred to as **SCL-EMG** in the remainder of this work.

4. GRADIENT DESCENT BASED CCC OPTIMIZATION

Since the task at hand is to create classifier predictions that perform well in terms of concordance correlation, it is natural to use this measure as optimization criterion. Solely optimizing for RMSE or MAE can lead to close trajectories, however the measure also punishes a low correlation coefficients and deviations of the variances of the signals. In the following, we derive the derivative of the CCC measure so that it can be used in stochastic gradient descent optimization of feed-forward multilayer perceptrons.

The concordance correlation coefficient is defined as follows

$$CCC(x,y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\bar{x}-\bar{y})^2} \tag{1}$$

and consists of different parts such as Pearson's correlation coefficient

$$\rho = \frac{\frac{1}{n}\sum_i^n (x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} \tag{2}$$

and the standard deviation and the mean of x and y. The derivative of the concordance correlation with respect to a input point x_k can be stated as follows.

$$\frac{\partial CCC}{\partial x_k} = \frac{\partial}{\partial x_k}\frac{2\frac{1}{n}\sum_i^n (x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}\frac{\cancel{\sigma_x}\cancel{\sigma_y}}{\sigma_x^2+\sigma_y^2+(\bar{x}-\bar{y})^2} := \frac{f}{g} \tag{3}$$

the individual parts of the derivative are as follows:

$$\frac{\partial f}{\partial x_k} : (y_k-\bar{y}) - \underbrace{\frac{1}{n}\sum_i (y_i-\bar{y})}_{0} \tag{4}$$

$$\frac{\partial g}{\partial x_k} : \frac{\partial}{\partial x_k}\Big[\underbrace{\frac{1}{n}\sum_i^n (x_i-[\frac{1}{n}\sum_j^n x_j])^2}_{(a)} + \underbrace{(\frac{1}{n}\sum_i^n x_i - \frac{1}{n}\sum_i^n y_i)^2}_{(b)}\Big] \tag{5}$$

$$\text{with } (a): \quad (x_k-[\frac{1}{n}\sum_j^n x_j]) - \underbrace{\frac{1}{n}\sum_i (x_i-[\frac{1}{n}\sum_j^n x_j])}_{0}$$

$$\text{and } (b): \qquad 2(\frac{1}{n}\sum_i^n x_i - \frac{1}{n}\sum_i^n y_i)\frac{1}{n}$$

The final derivative is the combination of equations 3 to 5.

$$\frac{\partial CCC}{\partial x_k} = \frac{\overbrace{\Big(y_k-\bar{y}\Big)}^{\frac{\partial f}{\partial x_k}}\overbrace{\Big(\sigma_x^2\sigma_y^2+(\bar{x}-\bar{y})^2\Big)}^{g}}{\Big(\sigma_x^2+\sigma_y^2+(\bar{x}-\bar{y})^2\Big)^2} - \frac{\overbrace{\Big(\frac{2}{n}(x_k-\bar{y})\Big)}^{\frac{\partial g}{\partial x_k}}\overbrace{\Big(\frac{2}{n}\sum_i^n (x_i-\bar{x})(y_i-\bar{y})\Big)}^{f}}{\Big(\sigma_x^2+\sigma_y^2+(\bar{x}-\bar{y})^2\Big)^2} \tag{6}$$

The derivative can now be used for gradient based optimization procedures such as stochastic gradient descent in feed forward multilayer perceptrons. Note that unlike more traditional loss functions like the L_2 loss for example, the gradient as described by 6 can only be used in batch mode. This originates from the necessary estimates of the mean and the variances of x. Since batch mode is normally used anyway because it exhibits a more stable convergence, this does not pose any practical restrictions.

The classifier used in this work is a feed-forward multilayer perceptron which is optimized on $1-CCC(x,y)$. It is equipped with latest techniques proposed by the deep learning community. We use mini-batches to comply with the loss function, dropout [24] as regularization, Adadelta [26] for optimization and, among others, parametric rectified linear units (PReLUs) [11] in the hidden layers. We construct a multilevel system using base classifiers and a combination scheme to leverage the multi-modality and make best use of the available data.

5. NUMERICAL EVALUATIONS

5.1 Experimental Setup

5.1.1 Preprocessing

Before the classification phase, a number of preprocessing steps have been carried out to reduce to computational complexity and to increase robustness of the trained models. Sub-sampling the data has two positive effects on the training procedure. First, by reducing the amount of samples, the training usually converges faster which is especially important in the development phase as many classifiers with different parameters and modalities have to be trained. Due to the large overlap between neighboring windows, the computed features are correlated. A reduction of the number of samples does therefore not impede the classification results. Secondly, different data subsets can easily be created using fixed offsets. This is helpful for ensemble classifiers that heavily rely on diverse input data.

Another preprocessing step that we used was to smooth the labels prior to classification. This is motivated by the feature computation. As most features are computed on windows with a length of several seconds, the time horizon spans more activity than can be found during a single frame of 40 ms duration. The label should therefore encode information over the same time span as well. We evaluated different methods, including max and min pooling and found that median and low pass filtering worked best.

5.1.2 Classification Phase

Based on the preprocessed data we constructed a number of different base classifiers, where two of them were taken from the literature, namely Breiman's Random Forest [3] using regression trees and Friedman's gradient boosting [8]. A leave one subject out cross-validation scheme was used to set the various parameters that are required by the learning algorithms as we empirically found that this procedure reflects the performance on the official test set better than the partitioning in the official train and validation sets.

For the CCC-NN the number of neurons was optimized together with the number of hidden layers in the networks. The optimization was done by randomly picking an architecture with 1 to 7 layers and up to 2000 neurons per layer. The transfer functions were also randomly sampled. For arousal an architecture was found with 4 hidden layers and 1580, 590, 890, and 1670 neurons for each layer.

For the Random Forest model a number of 250 individual regression trees was chosen, which is a compromise between computational costs and having enough trees to reflect the data. For this approach six individual models were created on the labels that were given by the six raters and a final result was obtained by averaging as it is done to obtain the ground truth.

Similarly, for the gradient boosting algorithm a total number of 5000 base learners was chosen. Further, in order to take advantage of the data 5 different runs of the approaches were made by using slightly different data samples that were selected by using a varying offset in the sub-sampling procedure. Hence a final result was obtained by averaging the 5 models.

The combination of the different features was conducted by frame-wise concatenation of the selected feature vectors. In order to select an optimal combination of modalities, a greedy feature channel selection was used. This procedure was carried out by first computing the performance of each feature individually, sorting the results and sequentially adding features channels to the input combination. The combination with the highest value was selected.

5.1.3 Post-processing

The prediction step of the regression models is followed by a number of post-processing techniques to improve the outcome. The first step was determined as a further smoothing step of the unprocessed predictions. This removes potential outliers and also reflects the continuity of the annotation as it is conducted using the label tool.

Another important post-processing procedure that helps to increase the accuracy of the predictions is to shift and scale the values of the predictions. This is conducted by setting the minimum and the maximum of the prediction for a sequence to the minimum and the maximum as it is found in the training set. The rest of the outputs are altered accordingly. One reason for this is the usage of the averaging in the ensemble methods that implicitly conduct some degree of attenuation of the deflections, which should be accounted for.

The final post-processing step is motivated by manual inspection of the predictions for the sequences, which showed that they seem to be shifted a little with respect to the true label. An additional shift forward in time was applied to compensate for this. Concretely a shift of 60 frames for both arousal and valence was applied.

5.2 Unimodal Results

We first evaluated the classifiers of choice on each modality separately. This was done to create an estimate of the discriminative power of the feature sets with respect to the concordance correlation. Table 1 summarizes the results of these experiments.

It is seen that features that are extracted from the audio signal render higher CCC values than the video signals for the affective dimension arousal. Contrary, the CCC values are higher for the video features compared to the ones based on the audio channel though on a generally lower level.

The comparison with the values reported in the baseline paper for the development set shows that the individual results for our approaches return higher CCC values except for three cases, i.e., for the official geometric features for the dimension arousal and the appearance based and EDA features for the dimension valence (see [19] for the precise values). To some extent, this is surprising as the authors of the baseline paper use a combination of different learning techniques and also optimize the respective meta parameters that they are testing with in this experiment.

Feature	**Gradient boosting**		**Random Forest**	
	arousal	valence	arousal	valence
Appearance*	0.293	0.308	0.313	0.313
Geometric*	0.236	**0.337**	0.172	**0.401**
HOG	0.236	0.282	0.200	0.250
PHOG-TOP	0.318	0.279	0.366	0.268
LBP-TOP	0.382	0.197	0.436	0.295
Audio*	0.383	0.135	**0.599**	0.199
LPC	0.532	0.100	0.549	0.130
MFCC	0.565	0.172	0.546	0.046
LFPC	**0.572**	0.134	0.549	0.087
ECG*	0.344	0.256	0.276	0.188
EDA*	0.125	0.236	0.110	0.148
ECG-PQRST	0.191	0.118	0.052	-0.016
ECG-Wavelet	0.194	0.103	0.063	0.017
SCL	0.084	0.231	-0.019	0.083
SCL-EMG	0.094	0.235	0.002	0.089

Table 1: CCC values for the individual modalities for the classifiers based on gradient boosting and random forest for the affective dimensions arousal and valence. The asterisk denotes the baseline features computed by the organizers.

5.3 Multi-modal Results

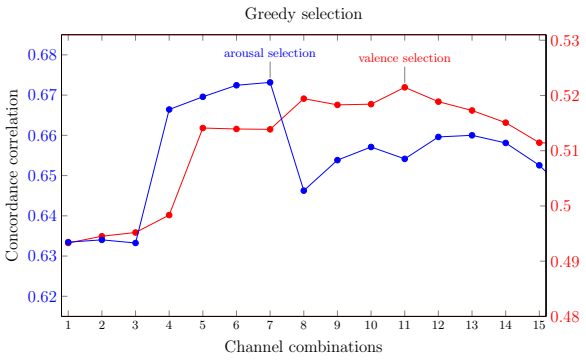


Figure 3: Greedy modality selection. By sorting the unimodal results a ranking is created which is used to obtain a suitable channel subset.

Based on the unimodal results, an optimal input channel subset was computed by subsequent greedy selection of the best modality each time. Figure 3 illustrates the resulting trajectories. For valence it shows a steady increase before the results start to become worse. In the arousal trajectory, a strong peak can be seen. We chose the highest points for both settings. Note, that while a greedy selection can be computed relatively quickly, it can only serve as a heuristic choice. The optimal combination can be something completely different. However how to obtain optimal input combinations is still an open research question.

Fusion scheme	Arousal	Valence
early fusion (RF)	0.508	0.409
early fusion (greedy, GB)	0.584	**0.448**
late fusion (NN on RF)	0.396	0.190
late fusion (NN on GB)	0.490	0.267
late fusion (CCC-NN on RF)	0.554	0.361
late fusion (CCC-NN on GB)	0.579	0.330
late fusion (fixed mapping)	**0.630**	0.381

Table 2: Multi-modal validation results (CCC). The greedy selection in combination with gradient boosting exhibits the highest performance. Late fusion using the CCC-NN is almost as good for arousal, but much worse for valence. Additionally we provide the results using a neural network optimized for RMSE using the same architecture. The CCC-NN outperforms it for both settings and both classifiers.

The selection process yields a combination of 7 channels for arousal and 11 channels for valence. The validation performance using the leave-one-subject-out procedure is illustrated in Table 2. As can be seen the selection dramatically improves the results for the validation for both arousal and valence.

While early fusion often performs quite well, in many cases fusion of individual classifier decisions using fixed or trainable mappings yields even better results. For this, a trainable fusion mapping using the introduced CCC-NN is conducted. The training set is divided into training sets for the base classifiers and the fusion mapping, respectively. The data was again split person-independently such that 9 subjects are used for the base classifiers (again Random Forest and Gradient Boosting) and 8 subjects for the fusion mapping. Testing was conducted on the left out subject. The averaged results over the 18 folds can again be seen in Table 2. A network with two tanh-layers and 49 neurons per layer serves as fusion mapping here. Late fusion using the CCC-NN is almost as good as the greedily selected early fusion with Gradient Boosting for arousal. For valence the results are worse than both early fusions. As a comparison we also provide results obtained from RMSE optimization of a neural network with the same architecture. The results show that the CCC-NN is better for both dimensions and both classifiers. Furthermore, we also conducted late fusion by averaging the results of the classifiers trained on the individual feature sets. Surprisingly, for arousal averaging yields the best validation results.

5.4 Test Set Submissions

We also evaluated our approaches on the official test set. For this, the classifiers were trained on the union of the train and validation set. For gradient boosting, 5 classifiers with 5000 base learners were trained with a relatively high subsampling factor of 300 with an offset of 60 frames between subsequent models. For the random forest again 250 trees were used. The results in Table 3 suggest that both models are clearly able to generalize on unseen subjects. Furthermore both submissions outperform the baseline that was provided by the challenge organizers. We also submitted a fixed late fusion, which again yielded the best results for arousal.

Setup	Arousal			Valence		
	RMSE	CC	CCC	RMSE	CC	CCC
Basel.	0.161	0.354	0.444	0.113	0.490	0.382
RF	0.187	0.526	0.520	0.139	0.453	0.449
GB	0.175	0.552	0.546	0.124	0.481	**0.479**
Fixed	0.159	0.687	**0.620**	0.138	0.490	0.478

Table 3: Results on the official test set.

6. DISCUSSION

6.1 Delay and Scaling

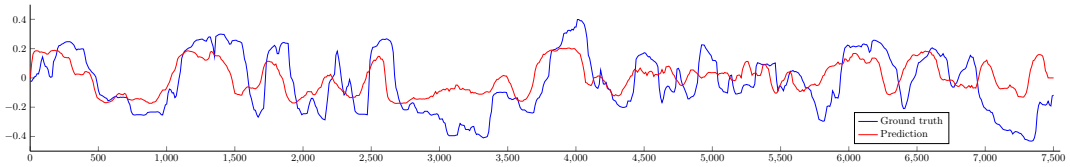
In this section we investigate the findings on the data set that are based on the post-processing steps outlined in Section 5.1.3. Figure 4 shows the effects that the post-processing has on the estimated trajectories and also in terms of CCC for arousal using early fusion with Gradient Boosting. Figure 4 (a) shows the raw classifier output. It can be seen that the range of the estimate is too small compared to the true labels. In Figure 4 (b) the minimum and maximum labels are set according to those in the training set. The performance thereby increases by 0.1 even though there is some degree of "overshooting" observable. When the estimated label is shifted by 60 frames to the right the CCC increases again by 0.2, resulting in a CCC of 0.748 as depicted in Figure 4 (c).

As mentioned before the scaling we used in our experiments defined by the minimum and maximum values found in the training set is not really optimal. An upper bound for this scaling approach is if the true maximum and minimum values would be provided by some oracle. We provide two little examples to illustrate this: For the dimension arousal when classifying the provided openEAR features using gradient boosting we get a raw CCC of 0.139, a value of 0.435 for the oracle and 0.383 for our approach in a subject independent cross-validation. Secondly regarding the dimension valence and HOG features with the same classification approach we get a CCC of 0.112 for the raw signal and 0.318 for the oracle scaling but only 0.282 for our approach.

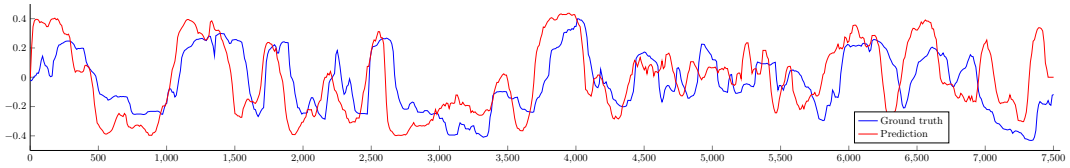
Considering the shifting in time of the prediction we additionally put forward one argument that it is more or less based on the used labeling strategy. Figure 5 shows the mean absolute label traces for the 6 raters over both dimensions for the first 250 frames of all sequences. It can be seen that there is an initial phase where the raters reach their "operational level" which lasts approximately 60 frames. The lower part of Figure 5 shows the performance for the label dimension arousal for the gradient boosting approach based on the whole proposed feature set. In concordance to the 60 frame long transient phase, the CCC has a maximum at also approximately 60 frames.

6.2 Protolabels Reloaded

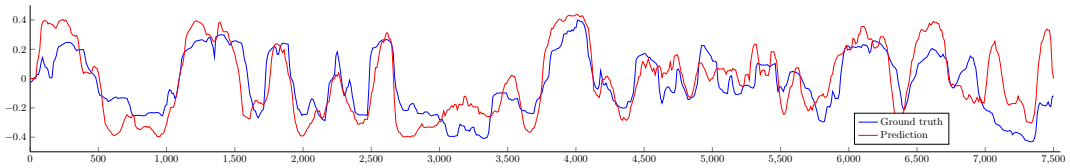
In 2014 we proposed protolabels for the classification of fully continuous affective states that are constructed from the training label information rather than from the actual audio or video data. This led to the first place of the affect sub-challenge, outperforming elaborate machine learning approaches such as deep neural networks and support vector machines. It was apparent that several issues were preventing the successful application of machine learning originating



(a) *Unprocessed* classifier output compared to the true label trace with CCC=0.456.



(b) *Scaled* classifier output compared to the true label trace with CCC=0.549.



(c) *Shifted* and *scaled* output compared to the true label trace with CCC=0.748.

Figure 4: Exemplary display of the effects of the post-processing steps for an arousal trace and the respective estimation.

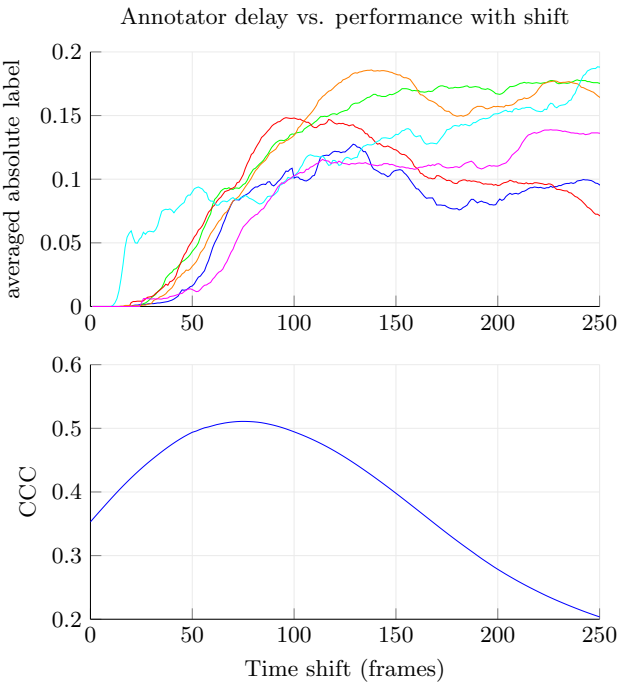


Figure 5: Annotator delay and classifier shifting: Mean absolute values for the 6 raters over all labels (top) and CCC performance related to the temporal shift of the classifier output.

from both the annotation procedure and the evaluation of performance. In the current edition of the challenge, it seems that the issues have been more or less resolved and that it is more fruitful to rely on data driven methods to classify the two affective dimensions (compare Figure 4). The main reason for that seems to be that the deflections of the label traces are large compared to the transient phase unlike in previous challenges. The exact reasons for this are open for speculation. It might be in parts due to the increased expressiveness of the data but maybe also because the chosen raters annotated more generously by assigning higher amplitudes. The spirit of the protolabels can however still be found as data independent steps can be carried out to improve the estimated trajectories. These operations are to scale the outputs to fit the range of the training set, to shift the label in time and to conduct several steps of label smoothing.

7. SUMMARY & CONCLUSION

In this work, we proposed a system for multi-modal affect recognition based on audio, visual and bio-physiological features. Results presented on the challenge dataset suggest that fusion of multiple modalities helps to improve the quality of the prediction over unimodal approaches. We highlighted the characteristic features of the new dataset and compared it with previous editions of the challenge. Furthermore, we presented insight into the necessary pre- and post processing steps that are common in machine learning applications but rarely discussed.

Possibilities for future work include, besides variations in classifier or feature selection, for example fusion with the help of uncertainty values (i.e. ensemble agreement, etc.). Another possibility is the use of techniques from the realm of partially supervised learning as the original RECOLA dataset contains additional video material of each participant that has not been annotated.

Acknowledgments

This paper is based on work done within the Transregional Collaborative Research Centre SFB/TRR 62 *Companion-Technology for Cognitive Technical Systems* funded by the German Research Foundation (DFG). Markus Kächele is supported by a scholarship of the Landesgraduiertenförderung Baden-Württemberg at Ulm University.

8. REFERENCES

[1] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):637–655, 1971.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of CIVR*, CIVR '07, pages 401–408. ACM, 2007.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Multi-scale temporal modeling for dimensional emotion recognition in video. In *Proceedings of AVEC*, pages 11–18. ACM, 2014.

[5] W. Chen, J. Zhuang, W. Yu, and Z. Wang. Measuring complexity using FuzzyEn, ApEn, and SampEn. *Medical Engineering & Physics*, 31(1):61–68, 2009.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893 vol. 1, 2005.

[7] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, Aug 1980.

[8] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[9] Z. Guoying and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[10] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of AVEC*, pages 33–40. ACM, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

[12] M. Kächele, M. Schels, S. Meudt, V. Kessler, M. Glodek, P. Thiam, S. Tschechne, G. Palm, and F. Schwenker. On annotation and evaluation of multi-modal corpora in affective human-computer interaction. In *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*, LNCS, pages 35–44. Springer, 2015.

[13] M. Kächele, M. Schels, and F. Schwenker. Inferring depression and affect from application dependent meta knowledge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 41–48. ACM, 2014.

[14] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm. Multimodal data fusion for person-independent, continuous estimation of pain intensity. In *Engineering Applications of Neural Networks*, Communications in Computer and Information Science, page (to appear). Springer International Publishing, 2015.

[15] M. Kächele, D. Zharkov, S. Meudt, and F. Schwenker. Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition. In *Proceedings of ICPR*, pages 803–808, 2014.

[16] H. Kaya, F. Çilli, and A. A. Salah. Ensemble cca for continuous emotion prediction. In *Proceedings of AVEC*, AVEC '14, pages 19–26. ACM, 2014.

[17] L. I. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989.

[18] T. L. Nwe, S. Foo, and L. De Silva. Classification of stress in speech using linear and nonlinear features. In *Proceedings of ICASSP.*, volume 2, pages II–9–12 vol.2, April 2003.

[19] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. The AV+EC 2015 multimodal affect recognition challenge: Bridging across audio, video, and physiological data. In *Proceedings of the 5rd ACM International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015.

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8, April 2013.

[21] M. Schels, M. Glodek, G. Palm, and F. Schwenker. Revisiting AVEC 2011 — an information fusion architecture. In *Computational Intelligence in Emotional or Affective Systems*, Smart Innovation, Systems and Technologies, pages 385–393. Springer, 2013.

[22] M. Schels, M. Kächele, M. Glodek, D. Hrabal, S. Walter, and F. Schwenker. Using unlabeled data to improve classification of emotional states in human computer interaction. *Journal on Multimodal User Interfaces*, 8(1):5–16, 2014.

[23] M. Schels and F. Schwenker. A multiple classifier system approach for facial expressions in image sequences utilizing gmm supervectors. In *Proceedings of the ICPR*, pages 4251–4254. IEEE, 2010.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[25] 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*, AVEC '14, pages 3–10. ACM, 2014.

[26] M. D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.