This is a copy of: Kächele, M. and Schels, M. and Thiam, P. and Schwenker, F.,
Fusion Mappings for Multimodal Affect Recognition, IEEE Symposium Series on Computational Intelligence, 2015, 307–313.
The final publication is available via http://dx.doi.org/10.1109/SSCI.2015.53

Fusion Mappings for Multimodal Affect Recognition

Markus Kächele*, Martin Schels*, Patrick Thiam and Friedhelm Schwenker
Institute of Neural Information Processing
Ulm University, 89069 Ulm
Germany
Email: markus.kaechele@uni-ulm.de

***Abstract*—Affect recognition is an inherently multimodal task that makes it appealing to investigate classifier combination approaches in real world scenarios. Thus a variety of different independent classifiers can be constructed from independent input modalities without having to rely on artificial feature views. In this paper we study a variety of fusion approaches based on a multitude of features that were extracted from audio, video and physiological signals for continuous recognition of spontaneous affect. For this purpose the RECOLA data collection is analysed. In uni- and multimodal experiments we show how an ensemble can outperform the best individual classifiers.**

I. Introduction

A. Ensemble learning and classifier fusion

Multiple classifier systems are an elegant and effective means to improve the classification accuracies of individual classifiers. The combination of an ensemble of classifiers succeeds if the individual classifiers show a distinct independence, which is commonly called diversity in this context [1].

In [2] three main approaches to create diverse base classifiers are outlined. The first approach is to use a different subset of training samples to create different classifiers. This approach is implemented by the popular ensemble techniques bagging [3] and boosting [4]. Second, using different feature representations for the same data points or subsets of the feature vectors for different models. A well established application of this principle is the random subspace method [5]. And finally using different learning techniques or initializations for the individual classifiers to achieve independent models.

An important issue in the course of classifier combination is the question whether to use an adaptive fusion layer or a fixed combination scheme [6]. Convenient fixed rule combiners are using decision voting for classifiers with crisp outputs [7] or summing or multiplying when fuzzy or probabilistic classification algorithms are used [8]. Using a trainable classifier fusion layer treats the task of combination of classifier outputs as a new classification problem that is stacked onto the initial one [9]. However, there is a need for further validation data for the training of the mapping. This reduces the amount of available data points for the training of the base classifiers [10].

B. Classifier fusion for emotion recognition

The classification of human affective states is an application that is particularly feasible to study multimodal fusion architectures. The conveyance of emotional signals relies on a multitude of independent channels, e.g., facial expression, non-verbal communication but also physiological signals that can be exploited to improve classification.

In [11] a fixed averaging rule combiner was used to classify individual feature views that were extracted from facial expressions. Further, in [12] a neural network based fusion architecture was proposed for the classification of three discrete affective dimensions in continuous time.

Sometimes blocked stimuli are used to induce affective states in test subjects. In [13] an automatic information fusion architecture was proposed to combine three independent modalities: audio, video and physiology (namely respiration, blood volume pulse and electromyography). This was conducted by assessing uncertainty values by means of a reject option for uncertain samples and subsequent multimodal and temporal integration.

Taking into account the temporal characteristics of the affective states a Markov chain based approach for the combination of multimodal and temporal classifier decisions, called "Markov fusion network", has been presented in [14]. In [15] the authors borrow the Kalman filtering concept, which has a long history in different areas of engineering, to model the uncertainty of the classification and also to combine multimodal decisions over time.

Finally, in [16] a proto-label based approach is proposed that incorporates multimodal properties for the construction of individual user groups. Despite it's simplicity, this approach achieved the best performance in the affect sub-challenge of the 2014 edition of the Audio/Video Emotion Challenge.

For an extensive overview over recent developments in multimodal classification of emotions the reader is referred to the comprehensive review papers of Zeng et al. [17] and Wu et al. [18].

The remainder of this work is structured as follows. In the next section, the data set and feature extraction pipeline are introduced. The utilized base classifiers are briefly introduced in Section III, followed by the fusion mappings, grouped into fixed and trainable in Section IV. Experimental validation is presented in Section V. The findings are discussed in Section VI followed by the conclusion.

* These authors contributed equally to this work.

978-1-4799-7560-0/15/$31

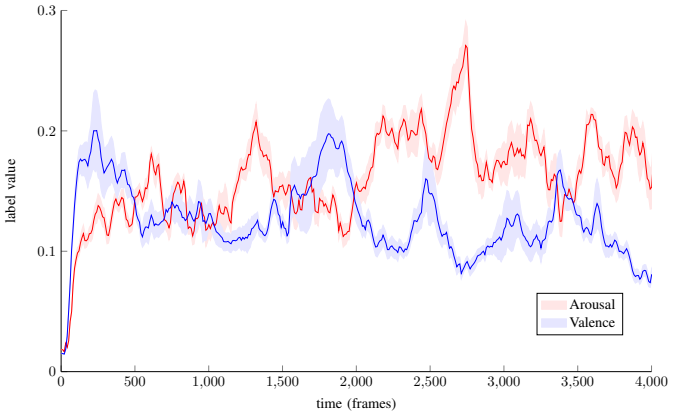


Fig. 2. Averaged traces for the provided labels "arousal" and "valence" over the 18 sessions. Their variances are also given as error corridors.

II. Data Set and Feature Extraction

The data collection that is used in this work is the RECOLA database that was recorded at the University of Fribourg, Switzerland [19]. It comprises 18 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 (compare Figure 2). Each recording was annotated by 6 native French speakers. The average of these individual ratings is used as ground truth. A sample screen shot of the data is shown in Figure 1.

Ringeval et al. [20] suggest to use Lin's concordance correlation coefficient (CCC) [21] for the evaluation of the classification performance on this dataset as it serves as combination of correlation and error based measures:

$$\rho_C = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\bar{x} - \bar{y})^2} \quad (1)$$

with x, y being a prediction and a true label, ρ is the correlation coefficient, σ^2 denotes the variances of the signals and $\bar{x}$, $\bar{y}$ are the means of x and y, respectively. The CCC can also take negative values and ranges from -1 to 1. It is a combination of Pearson's correlation coefficient and the RMSE of the two signals, where not only the linear relationship between the signals is measured as in other data sets [16] but also their respective shifts. Because a high correlation coefficient and a small deviation are necessary to achieve a high performance value, the measure overcomes some of the limitations of the individual terms that are outlined in [22].

In the following the features that were extracted from the data are briefly described.

A. Audio Features

Linear predictive coding coefficients (LPC) are extracted using an auto-regressive model. Predictions for a new sample are hereby created from the p preceding samples [23]. They are still widely used in fields such as speech recognition and speech synthesis because of their simplicity and ease of computation. Here, 8 coefficients were computed for time windows of 32 ms length with an offset of 16 ms.

Mel frequency cepstral coefficients (MFCC) [24] are one of the most popular feature descriptors for speech. They are obtained from the power spectrum of a short time Fourier transform, followed by conversion to the mel scale by $M(f) = 2595 \log(1 + f/100)$ (for a frequency f in Hz) and triangular bandpass filtering. The discrete cosine transform is then applied to quantise the signal to the desired number of coefficients.

We used 20 MFCC coefficients per time window of length 32 ms with an offset of 16 ms for the experiments here.

Similarly to MFCC, the **log frequency power coefficients** [25] are computed from the power spectrum of windowed speech segments. A log filterbank is applied that is inspired by the frequency resolution of the human ear ranging from 100 Hz to the Nyquist limit [25].

Furthermore we used the openSMILE toolkit [26] to extract additional features. 42 additional low level descriptors from energy, spectral and voicing related feature groups are computed on short time scales and then integrated using statistical moments.

B. Video Features

Before computing the video features, the facial region was detected and aligned using stable landmarks obtained by the Supervised Descent Method [27]. Alignment is necessary to compute meaningful features that span more than one frame and to consistently capture the same subregions. On the aligned faces, the following features have been computed.

Local binary patterns in three orthogonal planes (LBP-TOP) [28] are computed on image sequences and thus encode information about the temporal order of the input signals. In comparison to volumetric local binary patterns, the LBP-TOP descriptor samples the space-time volume only on three predetermined slices, arranged orthogonally along the $x - y$, $x - t$ and $y - t$ planes.

The LBP-TOP coefficients are computed using a window length of a second. To encode additional spatial information, the facial region is subdivived into 2×2 blocks that are overlapping by 25%. The final descriptor is assembled by concatenation of the blockwise computed LBP-TOP descriptors.

Local Gabor binary patterns from three orthogonal planes (LGBP-TOP) [29] are similar to LBP-TOP. The main difference is that before computing the LBP descriptors on the orthogonal planes, the images are filtered using a Gabor filter bank with different scales and orientations. Gabor filters are inspired by the human visual system and are commonly used to capture oriented structure. Since the resulting feature is quite high dimensional, principal component analysis is applied for dimensionality reduction.

Pyramids of histograms of oriented gradients in three orthogonal planes (PHOG-TOP) is another instance of an orthogonal plane descriptor. Instead of LBP however, PHOG [30] is computed for the different planes. PHOG combines spatial information with the distribution of image gradient orientations using a pyramidal scheme 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

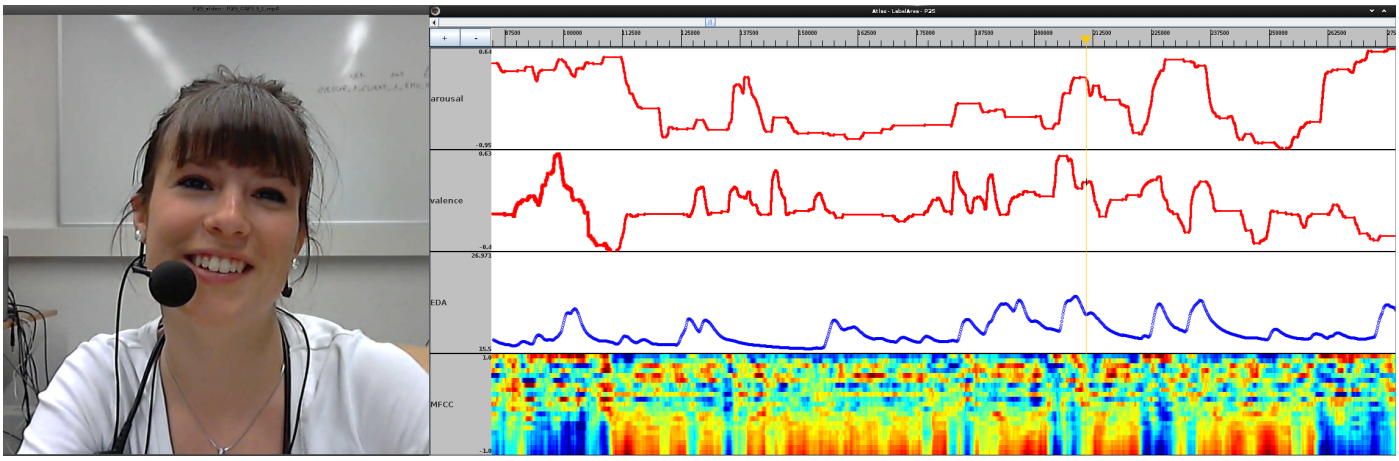


Fig. 1. Exemplary recording situation of the RECOLA data set. The figure shows a screenshot of the video, a sample audio signal represented as MFCC coefficients, an electrodermal activity sample curve and the traces for the two label dimensions "valence" and "arousal".

cell, a HOG descriptor is computed. The final PHOG descriptor emerges as a concatenation of all HOG descriptors over every pyramid level.

Another descriptor for localized structures in images are the **histograms of oriented gradients (HOG)** [31]. First, gradients of salient edges in the image are computed which serve as input for a histogram of gradient directions with n orientation bins.

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 which renders a descriptor dimensionality of 324.

As an additional feature, landmark distances as computed in [20] are used as well (termed **geometric** in Section V).

C. Bio-Physiology

Emotional reactions can be inferred from bio-physiological measurements such as the electrodermal activity (EDA), electrocardiogram (ECG) or electromyography (EMG) since they are directly controlled by the autonomic nervous system [32].

For the recorded **ECG** channel, the following features have been computed. Heart rate variability, zero-crossing rate, non-stationarity index, spectral entropy, based on a power spectrum density: low and high frequency powers as well as their ratio. Additionally statistical moments and derivatives have been computed on the signal.

In the **EDA** channel, responses to external stimuli are reflected in a change of skin conductivity, which forms activity peaks over time and is called the phasic component. The other part of the signal is the tonic component which can be regarded as a baseline that slowly drifts over time. For the two components, the spectral entropy, mean frequency, non-stationary index are computed. Additionally, the first coefficient of a regression model and statistical moments are computed as well as the first derivative.

III. Base Classifiers

As base classifiers we decided to use Random Forests [33] and gradient boosting [34]. Note that a multitude of different choices is possible here. We decided for those classifiers because they are robust and work relatively well without excessive need of meta parameter tuning.

The Random Forest [33] is an ensemble of bagged decision or regression trees. Each tree is computed on a bootstrapped feature subset to maintain diversity. For regression, the result is a subdivision of the input space into axis aligned cells that each represent a continuous value.

Gradient boosting [34] is a variant of boosting that incrementally adds new weak learners based on the negative gradient of the selected loss function in each optimization step.

IV. Multiple Classifier Fusion

In this section, the different combination schemes are presented. They are grouped into fixed and trainable combiners. Fixed combiners take the base classifier predictions as input and apply a fixed mapping to it. They have the advantage that no additional training phase occurs and therefore all the data can be used for optimizing the base learners.

Trainable mappings on the other hand are trained on the predictions of the base learners. Using a trainable mapping, systematic confusions can more easily be resolved than with a fixed mapping. The downside is that additional data is necessary to train the combiner. This can be especially critical in scenarios where hardly enough data exists to properly train the base learners. The trainable mappings additionally bring the complication of how to optimally split the data into training sets for the base learners and the fusion mapping. For solutions regarding this problem, the reader is referred to [10].

In the following, C^i will denote the output of classifier i and y will denote the ground-truth labels. Furthermore

$$\mathbf{C} = \begin{bmatrix} | & & | & & | \\ C^1 & \cdots & C^i & \cdots & C^n \\ | & & | & & | \end{bmatrix}$$

is the concatenation of the individual classifier predictions in matrix form.

A. Fixed Rule Combiner

The simplest approach to combine individual classifier decisions is to apply a fixed combination rule to the outputs of the classifier [7]. One intuitive choice for such a fixed rule combiner that is applicable for the classification of continuous dimensions is to take the average of the individual models [8]. Thus, the individual errors that are committed by the base models are averaged out if the models are independently created. Further, we evaluate the median rule as a small variation of this concept, which is more robust against outliers.

B. Trainable Fusion

1) Model free approaches: Model free combiners are created by assessing the performance of the individual models and then combining their outputs accordingly. We evaluate two different techniques of this category namely computing a weighted average of the outputs and choosing the single best model. In order to obtain suitable weights for the weighted average, each classifier is tested on the validation set

- Weighted average: The individual models are evaluated on the validation set and the weights of the combination are set with respect to this performance.
- Single best: The modality is picked that achieved the highest performance given the predictions on the hold-out set.

2) (Regularized) linear Optimization:

- Pseudo-Inverse: A least-squares optimal linear mapping is obtained by computing the pseudo-inverse of the classifier outputs $\mathbf{C}$ and multiplying it with the desired values y.

$$M = \lim_{\alpha \to 0_+} \mathbf{C}^T(\mathbf{C}\mathbf{C}^T + \alpha I)^{-1} y \tag{2}$$

  The mapping is then applied to the predicted outputs to obtain the final class memberships. For details, the reader is referred to [35].

- Robust regression: It has been proposed as a more robust alternative of ordinary least squares which can be very sensitive against outliers in the data. The solution is obtained by iteratively reweighted least squares. This is an iterative process in which a weighted least squares solution is computed, followed by weight updating of points based on their distance to the new model. Here, the Huber weighting function is used:

$$h(x) = \begin{cases} \frac{1}{2}x^2 & |x| < t, \\ t(|x| - \frac{1}{2}t) & \text{otherwise} \end{cases} \tag{3}$$

  with a threshold value t.

- Elastic net: A regression formula with a convex weighting of ℓ_1 and ℓ_2 regularization. For $\alpha = 1$, the LASSO [36] is obtained while the formula simplifies to ridge regression for $\alpha = 0$ [37].

$$\min_{\beta} \|y - \mathbf{C}\beta\|^2 + \lambda \left(\alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1\right) \tag{4}$$

  Values in between trade off properties of both algorithms. The ℓ_1 loss favours sparse models, but has problems with correlated variables, which can be handled more efficiently by the ℓ_2 loss.

3) Classifiers:

- Random Forest: See Section III.
- ϵ-SVR: A regression variant of the well known Support Vector Machine. The optimization criteria is the ϵ-insensitive loss. That means that predictions are desired that are closer than ϵ from the data points; larger deviations are penalised. As it is the case for the SVM, the kernel trick can be used to extend the capabilities of the SVR to nonlinear problems. Here we use SVR with linear and RBF kernel.
- CCC-Net: A feedforward multilayer perceptron optimized directly on the concordance correlation coefficient. Stochastic gradient descent is used along with the derivative of Eq. 1 for the optimization task. A network with two hidden sigmoid layers and 49 neurons per layer is used as combination mechanism.

V. Numerical Evaluations

This section is divided into uni- and multimodal results. The unimodal results lay the foundation on which the multimodal fusion builds. The experiments are designed to give an estimate on the person independent generalization ability of the classifiers using a leave-one-subject-out cross validation procedure. In case a trainable classifier fusion approach is stacked on the individual models a validation set is separated from the training set by randomly choosing 9 subjects for the training of the base classifier with the remaining 8 subjects being used to construct the combiner.

Preliminary experiments showed that it is mandatory to conduct distinct pre-processing and post-processing steps in order to obtain a good performance. The complete work-flow of the experiments comprising classification and fusion but also pre- and post-processing is shown in Figure 3. Before the classification is conducted, the data is sub-sampled by a factor of 20 in order to make the training of the base classifiers computationally feasible. Further the training labels are smoothed using a median filter with a window size of 200 frames. The low frequency of the label traces together with the long time windows with large overlap on which the features have been extracted ensure that the mentioned steps do not deteriorate the prediction quality. Also, since the extracted features span a longer time window it seems implausible to assign single label values that span only 40 ms to them.

After the base classifiers and the fusion step were applied to a sample two main post-processing steps are carried out: The output is scaled such that the minimum and maximum value equal the minimum and maximum values as they are found in the training set. This corrects the attenuation that a signal implicitly undergoes when applying an ensemble method such as the Random Forest but also some of the trainable combiners we applied for multimodal fusion. As a second step the signal is shifted by 60 frames. This procedure increases the CCC considerably. It might be caused by the delay in the labeling procedure as described in [19]. Another possible reason is the extraction of features based on larger time windows that makes

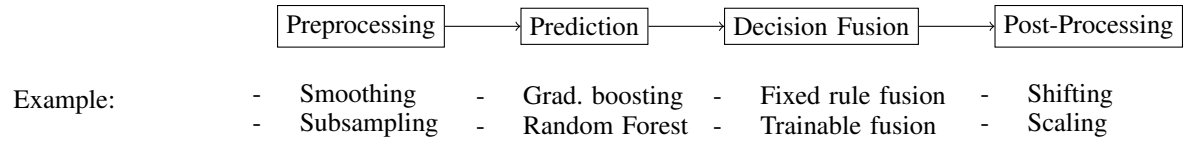


Fig. 3. The information processing comprises pre-processing, classification, fusion and post-processing.

Feature	Gradient boosting		Random Forest	
	arousal	valence	arousal	valence
LGBP-TOP	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
openSMILE	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

TABLE I. CCC VALUES FOR THE INDIVIDUAL MODALITIES WITH THE BASE CLASSIFIERS GRADIENT BOOSTING AND RANDOM FOREST FOR THE AFFECTIVE DIMENSIONS AROUSAL AND VALENCE.

Fusion mapping	Random Forest		Gradient boosting	
	arousal	valence	arousal	valence
Mean	**0.630**	**0.381**	0.539	0.345
Median	0.557	0.325	0.503	**0.350**
Single best	0.423	0.380	0.296	0.196
Weighted average	0.627	0.350	0.534	0.336
Pseudo inverse	0.571	0.369	0.505	0.320
Robust regression	0.574	0.349	0.503	0.307
LASSO	0.579	0.374	0.503	0.319
Elastic net	0.580	0.369	0.505	0.318
ϵ-SVR (linear)	0.570	0.363	0.506	0.315
ϵ-SVR (RBF)	0.541	0.226	0.455	0.231
Random Forest	0.521	0.363	0.424	0.316
CCC-NN	0.554	0.361	**0.579**	0.330

TABLE II. CCC VALUES FOR THE COMBINED MODALITIES USING FIXED AND TRAINABLE MAPPINGS.

it in a sense unclear what the real value for a feature vector is compared to those in a vicinity.

A. Unimodal results

The unimodal results were computed by training and testing on a single modality. The results for Random Forest and gradient boosting can be seen in Table I.

For arousal, audio features like MFCC, LPC and LFPC show high performances for both classifiers. While the openSMILE features exhibit only a moderate performance with the Random Forest, they achieved they highest unimodal result of 0.599 using gradient boosting. For valence video features like the landmark distances and LGBP-TOP are the most discriminative ones, followed by bio-potentials classified with the Random Forest (ECG: 0.256 and EDA: 0.236).

The fact that the employed classifiers often result in different predictions and thus different performances for the same features can be seen as a foundation for multimodal experiments. It suggests that by the different training algorithms (i.e. decision tree building vs. boosting using the gradient of the loss function) other optima are achieved that behave unlike each other given unseen examples.

B. Combined Results

For the multimodal results, the combination techniques presented in Section IV are applied to the unimodal predictions. For the fixed mappings, the base classifiers are trained on the whole dataset, while for the trainable mappings, the mentioned split into a training (9 subjects) and a validation set (8 subjects) is conducted. The results are summarized in Table II.

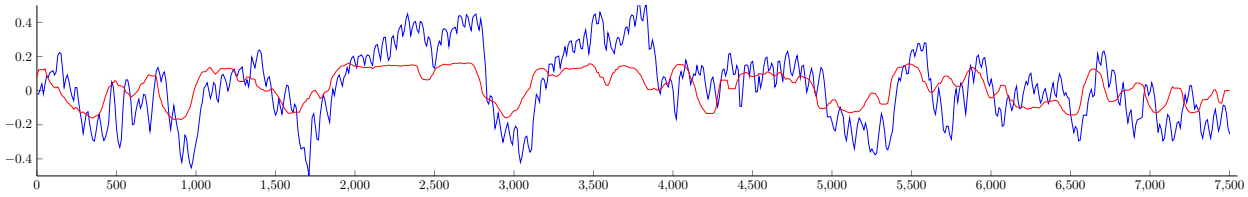
The table shows that fixed rule combiners are comparably robust with the averaging combiner achieving the highest accuracy for both affective dimensions. One reason for that is that they have the whole training set at their disposal and hence can build on more accurate base classifiers. This effect can be observed in the row for the single best model selection which tends to be the worst performing combination method. The negative effects of the reduced training data can, however, be alleviated by combining all available models with respect to their performance on the validation set. Applying different types of linear optimization techniques from plain pseudo inverse to linear ϵ-SVR yield almost identical CCC values for all types of experiments. By conducting non-linear fusion, the additional degrees of freedom cannot be exploited to get better results and are prone to over-fitting in this case. One exception to this is the CCC-NN that renders a promising result especially for the dimension arousal and gradient boosting as base learner.

To summarize, the fixed combiners (especially the mean) show excellent performance with 0.630 for arousal and 0.381 for valence when classified with Random Forest. For gradient boosting, the median combination exhibits the best performance for valence while the CCC-NN performs best for arousal.

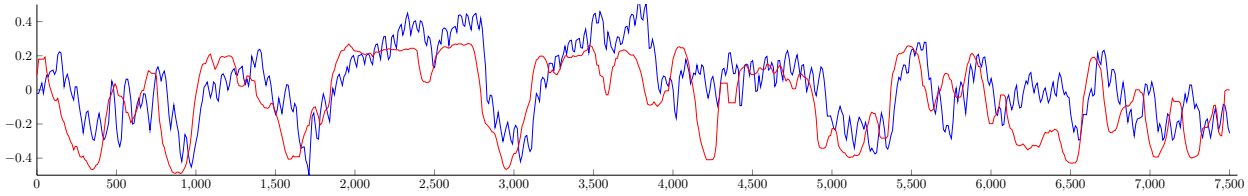
VI. DISCUSSION

A. Delay and scaling

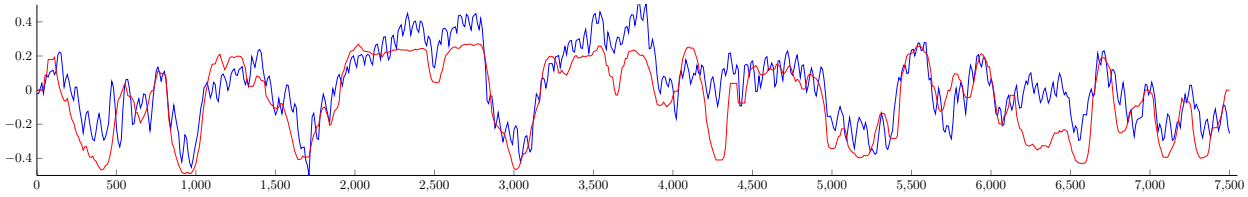
During the experimentation process we noticed, that the predicted traces that resulted from the combination process are condensed in the middle of the value range (around 0). That means that while the trajectory might be similar, the amplitude of the original signal is never quite reached (see Fig. 4 (a) for an illustration of this). Therefore, we decided to apply a fixed scaling to get back to the range of the original training labels. Note, that no information of the test labels is used. Furthermore, we noticed that the predictions have a small time delay of about one to two seconds. We believe that this is caused by the annotation delay of the individual raters. In Figure 5, this issue is illustrated. The per-annotator averaged absolute label traces of the beginning of the recordings are



(a) *Unprocessed* classifier output compared to the true label trace with CCC=0.481.



(b) *Scaled* classifier output compared to the true label trace with CCC=0.587.



(c) *Shifted* and scaled output compared to the true label trace with CCC=0.763.

Fig. 4. Exemplary display of the effects of the post-processing steps for an arousal trace (blue) and the respective estimation (red).

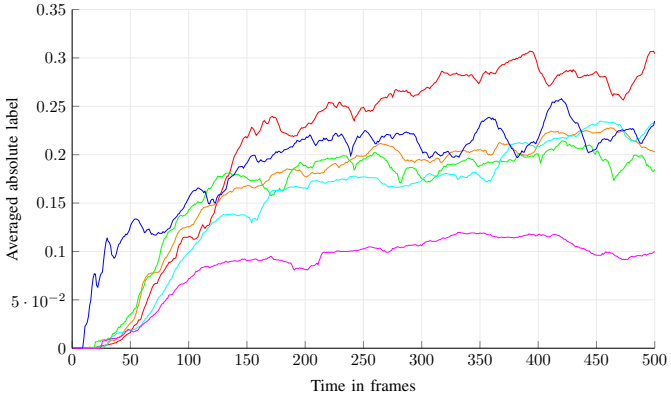


Fig. 5. Averaged absolute label traces for the 6 individual raters. An initial transient phase is clearly visible for each trace.

shown. It can be seen that there is an initial phase where the raters reach their "operational level" which lasts approximately 60 frames. To compensate this effect, a fixed shifting of the predictions by -60 frames is conducted. Note that those values are set globally and they do not depend on the testing set. Furthermore cross validation based investigations on an extra hold out set support this procedure.

To see the effect of the scaling and shifting procedures, the reader is referred to Figure 4 in which the impact of each step is visualized.

Figure 4 (a) shows the raw classifier output and it can be clearly 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 almost 0.2, resulting in a CCC of 0.763 as it can be seen in Figure 4 (c).

VII. Summary and Conclusion

In this paper we proposed and evaluated different classifier fusion strategies for the classification of multimodal data into continuous affective dimensions. A variety of trainable and fixed rule combiners are used to combine outputs of models that were constructed on 11 independently created features. As trainable combiners we used both, linear and non-linear learning approaches.

The main finding of the conducted experiments is that, here, the averaging combiner is the most accurate and stable approach. As mentioned before one reason for this is that no data is omitted from the training process for the base classifiers, which allows more accurate individual models. However, the CCC-NN is an interesting and promising approach in one of the cases we considered, especially when the accuracy of the individual classifiers is comparably low.

One further approach that may be apt to improve trainable fusion mappings is to conduct a modality selection step before the actual fusion is applied. Further, in order to construct a higher amount of diverse individual classifiers more heterogeneous base classifier approaches could be incorporated into the architecture.

Acknowledgment

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.

REFERENCES

[1] L. I. Kuncheva, "That elusive diversity in classifier ensembles," in *Pattern Recognition and Image Analysis*, ser. Lecture Notes in Computer Science, F. Perales, A. Campilho, N. Blanca, and A. Sanfeliu, Eds. Springer, 2003, vol. 2652, pp. 1126–1138.

[2] S. Pal and A. Pal, Eds., *Combining classifiers: Soft computing solutions*. World Scientific Publishing Co., Singapore, 2001, pp. 427–452.

[3] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[4] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[5] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[6] R. P. W. Duin, "The combining classifier: to train or not to train?" in *16th International Conference on Pattern Recognition*, vol. 2. IEEE, 2002, pp. 765–770.

[7] L. Kuncheva, *Combining pattern classifiers: Methods and Algorithms*. Wiley, 2004.

[8] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[9] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[10] C. Dietrich, G. Palm, and F. Schwenker, "Decision templates for the classification of bioacoustic time series," *Information Fusion*, vol. 4, no. 2, pp. 101 – 109, 2003.

[11] M. Schels and F. Schwenker, "A multiple classifier system approach for facial expressions in image sequences utilizing gmm supervectors," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*, A. Ercil, Ed. IEEE, 2010, pp. 4251–4254.

[12] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction (ACII'11) - Part II*, ser. LNCS 6975, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer, 2011, pp. 359–368.

[13] M. Schels, M. Glodek, S. Meudt, M. Schmidt, D. Hrabal, R. Böck, S. Walter, and F. Schwenker, "Multi-modal classifier-fusion for the classification of emotional states in woz scenarios," in *1st International Conference on Affective and Pleasurable Design*. USA Publishing, 2012, pp. 5337–5346.

[14] M. Glodek, M. Schels, G. Palm, and F. Schwenker, "Multiple classifier combination using reject options and Markov fusion networks," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. ACM, 2012, pp. 465–472.

[15] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," in *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, M. De Marsico, A. Tabbone, and A. Fred, Eds. SciTePress, 2014, pp. 671–678.

[16] 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*, ser. AVEC '14. ACM, 2014, pp. 41–48.

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

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

[19] 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*, April 2013, pp. 1–8.

[20] 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.

[21] L. I. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[22] 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*, ser. Lecture Notes in Computer Science, R. Böck, F. Bonin, N. Campbell, and R. Poppe, Eds. Springer International Publishing, 2015, pp. 35–44.

[23] 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*, vol. 50, no. 2, pp. 637–655, 1971.

[24] 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*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[25] T. L. Nwe, S. Foo, and L. De Silva, "Classification of stress in speech using linear and nonlinear features," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 2, April 2003, pp. II–9–12 vol.2.

[26] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. ACM, 2013, pp. 835–838.

[27] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 532–539.

[28] 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*, vol. 29, no. 6, pp. 915–928, 2007.

[29] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, ser. ACII '13. IEEE Computer Society, 2013, pp. 356–361.

[30] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, ser. CIVR '07. ACM, 2007, pp. 401–408.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, 2005, pp. 886–893 vol. 1.

[32] 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*, vol. 8, no. 1, pp. 5–16, 2014.

[33] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

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

[35] F. Schwenker, C. R. Dietrich, C. Thiel, and G. Palm, "Learning of decision fusion mappings for pattern recognition," *International Journal on Artificial Intelligence and Machine Learning (AIML)*, vol. 6, pp. 17–21, 2006.

[36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[37] H. Zou and T. Hastie, "Regularization and variable selection via the

elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.