Inferring Depression and Affect from Application Dependent Meta Knowledge

Markus Kächele
Institute of Neural Information Processing
89069 Ulm
Germany
markus.kaechele@uni-ulm.de

Martin Schels
Institute of Neural Information Processing
89069 Ulm
Germany
martin.schels@uni-ulm.de

Friedhelm Schwenker
Institute of Neural Information Processing
89069 Ulm
Germany
friedhelm.schwenker@uni-ulm.de

ABSTRACT

This paper outlines our contribution to the 2014 edition of the AVEC competition. It comprises classification results and considerations for both the continuous affect recognition sub-challenge and also the depression recognition sub-challenge. Rather than relying on statistical features that are normally extracted from the raw audio-visual data we propose an approach based on abstract meta information about individual subjects and also prototypical task and label dependent templates to infer the respective emotional states. The results of the approach that were submitted to both parts of the challenge significantly outperformed the baseline approaches. Further, we elaborate on several issues about the labeling of affective corpora and the choice of appropriate performance measures.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Applications; H.1.2 [**User/ Machine Systems**]: [Human factors, Software psychology]

General Terms

AVEC 2014

Keywords

AVEC 2014; affect recognition; depression recognition; meta knowledge

1. INTRODUCTION

Human-computer interfaces that go beyond the normal question-answer mechanism have received increasing attention in computer science. An interesting new channel for these interfaces is the automatic recognition of human dispositions [35, 2, 48]. The AVEC challenge has been established as a source of benchmarking data collections and also for the evaluation of classifier and machine learning approaches [51, 58]. This enables the development of various novel classification and regression approaches that are able to succeed in this challenging application [7, 11, 42, 19, 8, 20, 23, 53].

This paper presents our contribution to the fourth edition of the AVEC challenge. We argue that the consideration of background knowledge, meta-data about the subject and distinct knowledge about the respective application are the key factors for the recognition of target labels for the provided data set, where the affective expressiveness of the test subjects is low and the variance between the subjects is high. This entails the circumstances of the recording of the data set as well as the procedure of its annotation with appropriate labels.

The paper is structured as follows: in section 2, we will discuss the applications of affective human-computer interaction and the construction of respective corpora. This leads to the automatic classification and regression approaches, that are described and also numerically evaluated in section 3. Finally, section 4 offers some conclusions.

AVEC'14, November 7, 2014, Orlando, FL, USA.

ACM 978-1-4503-3119-7/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661806.2661813.

2. GENERAL CONSIDERATIONS

The automatic recognition of human affective states is a very challenging task [19]. This holds particularly true, when not only acted emotional patterns are considered, but also realistic expressions as they might as well occur in everyday live or during an interaction with a technical system [17].

The different branches of research in affective computing and related fields usually comprise three main steps: the design of (multi-modal) affective corpora, the creation of reliable ground truth labels (for example by self-rating or remote annotation), and the analysis of the recorded data. This analysis step can be conducted in various ways. The spectrum ranges from statistical evaluation of (self-assessed) categorical or meta data (e.g., from questionnaires) to highly sophisticated machine learning algorithms leveraging multimodal sensory input (e.g., audio, video, physiology) to create robust estimators using complex architectures comprising feature calculation, selectively trained classifiers and information fusion schemes [40, 48].

2.1 Affective Computing

Affective states and dispositions can be inferred from a broad spectrum of sensory input. One of the most commonly used signal domains is the audio channel and hence

human speech including its content and prosodic characteristics [47, 4]. For the recognition of emotions from speech, various approaches have been introduced. A main topic of research is the computation of meaningful features. Different feature sets have since been introduced such as linear predictive coding (LPC) [33], PLP [13] and RASTA-PLP [14]. More recently, it has been found that for the task of emotion recognition features such as MFCC [36, 32, 31] and LFPC [34, 24] seem to outperform other choices.

Additionally to the analysis of audio signals, many different approaches have been followed on emotion recognition from visual input for example in the form of facial expressions, movement cues [57] and body gestures [5]. Further, significant efforts have been made to develop feasible feature descriptors, localization and robust tracking of facial features and classification and information fusion techniques based on one or more feature sets [22, 50].

Another interesting channel for the detection of affective states is bio-physiology including heart rate, skin conductivity, electromyography (EMG), electroencephalography (EEG) and breathing rate. The reason for this is that the activity of the autonomous nervous system is directly correlated with emotional response, especially with respect to the arousal dimension in the VAD space [39], which can be inferred more or less directly from body measurements [25, 44, 21, 45].

Reliable affect recognition in each of those modalities is however still an ongoing topic of research and far from being solved [60, 25, 61]. As with each sensor input's characteristics there are also drawbacks that impede successful predictions and have to be kept in mind when dealing with a particular input modality. Audio based recognition for example is unavailing if there is no detectable speech [31]. Video based recognition heavily depends on the successful detection of the face. Intense movements and rotations complicate this step and may lead to false detections and therefore unsuitable inputs for predictions [42]. Bio-physiological measurements are also prone to movement artifacts (because of the physically attached sensors) and have the additional drawback that often a long time window is needed (>30 seconds) to make reliable estimations [40]. One possible way to circumvent those problems is to make predictions based on multi-modal input signals.

Multi-modal recognition of affective states most commonly deals with audio visual data, as it is also the case for the AVEC 2014 competition. The trend however goes towards an additional incorporation of bio-physiology, either solely combined with video, as found for example in the DEAP corpus [27] or combined with both audio and video, as e.g. found in the EmoRec II [59], Recola [37] or HCI Tagging [55] corpora.

In real world applications, systems based on multiple modalities generally outperform systems based on a single modality only [40, 19]. The question that remains to be answered is how to combine predictions of different modalities so that an actual improvement is achieved [48, 41, 10]. This question is not easily answerable and there are research fields that focus solely on the combination of information from multiple sources [28, 26]. Various ways exist to combine different input modalities. One possibility is to directly concatenate features of different modalities before training a classifier (this process is known as early fusion).

Analogously, fusion can be carried out after individual classification results are available (late fusion) for example

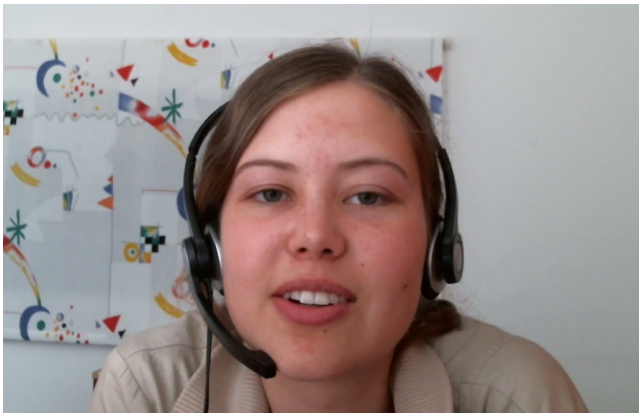


Figure 1: Exemplary recording situation for the AVEC data set. Audio-visual data is collected using a retail webcam and a headset in an unconstrained inquiry-response cycle.

static aggregation rules (e.g., sum or product rule), or by training of an additional classifier on top [28, 52]. More sophisticated fusion methods exist that exploit additional characteristics such as time continuity, classifier confidence, and sampling rates. Algorithms that fall in this category are the Markov Fusion Network (MFN) [7, 9], which allows to assign confidence measures for each classifier decision to be combined using the so called data potential or the Kalman filter [6], which infers the decision for a new sample based on a model estimated from already seen data and a measurement step which also contributes to the model update.

For the fusion of multi-modal signals in time continuous space, the use of recurrent neural networks has also become appealing. Their dynamic nature allows the exploitation of nonlinear time dependencies between feature vectors of different modalities in a sequence. Popular choices include recurrent neural networks that are not affected by the vanishing gradient problem [15] which can be trained over longer time periods such as Long Short-Term Memory Networks (LSTM) [16] and Echo State Networks (ESN) [18]. Both approaches have successfully been employed for the fusion of audio visual signals. In [61] an LSTM network has been used in the context of affect recognition while in [49] ESN have been employed for the audio-visual detection of laughter in conversations.

In social signals, emotional events often occur only rarely or without clear attribution to one of the given classes. Hence there is usually a large amount of recorded material that does not directly influence the training of classifiers. It has been shown that despite the lack of reliable label information, additional information in this form can be highly profitable for classification algorithms that leverage techniques of *semi-supervised learning (SSL)* [54] for their training. In [43, 46] for example, the authors applied an unsupervised preprocessing step using unlabeled data to transform the labeled samples into another representation which improves the classification rates. Other examples for SSL include semi-supervised annotation of corpora using Co-training with Tri-Class SVMs [12] or combinations of active learning and self-training [4].

2.2 Annotation

A further non-trivial issue for the assembly of affective corpora is the annotation of the recorded materials with labels that reflect a user's state adequately. There are mainly three different approaches that are followed to assign labels in this application. They all try to circumvent the fact that the true state is commonly unknown and also not exactly assessable.

The most straightforward approach to determine the affective state is to query it from the respective subject. For this purpose different questionnaires and pictographic techniques have been developed to infer emotional states. It is however not really possible to reflect short or medium term changes of affective states using this technique as it is desirable for human-computer interaction scenarios.

One rather popular method is to design different external stimuli that are presented to the test subject in order to elicit a desired affective state [60, 38]. This comprises often different difficulties of a given task or making the interaction with the technical system more difficult, e.g., by impairing the reactions to the given commands. Hence, the different target states are one after the other processed in a carefully designed experimental protocol.

An alternative approach is to manually label the material after the recording step [58]. As mentioned before, it is very difficult to infer an emotional state of a subject from the outside when the affective display is only subtle. Hence a large number of raters is required in order to average out the errors that the individual raters are assumed to commit. A further issue is that there is not really a convenient procedure for the annotation process. Using a continuous annotation method where a label value is manipulated in real time might lead to a comparably fast annotation process but might suffer from the individual attentiveness and the reaction times of the respective rater. The annotation of categorical labels is even more complex and commonly requires a large amount of navigating in the assigned material [30, 48].

2.3 Corpora

Notable multi-modal data collections that have been constructed in the application of human-computer interaction are outlined in the following. One general approach is to instruct a test subject to solve a specific task using a computer. An Example for this kind of data collection is the EmoRec II corpus, where a subject is playing multiple rounds of a card game using a voice controlled dialog system [59]. Different user states are elicited by giving positive or negative feedback to the subject by using different difficulty levels for the game. The corpus comprises a variety of different modalities with audio and video data but also physiological recordings. Another example for a task driven approach is the last minute corpus where a subject [38] is cooperatively interacting with a dialog system. Different user states are hence induced by a malfunctioning user interface or the constraints, that are imposed by the task.

A quite contrary approach is to allow an interaction with a computer that is as unconstrained as possible. One example for such a data set is the sensitive artificial listener data set, that has been used for the first two editions of the AVEC competition [51]. The test subject is situated in front of a computer screen displaying an artificial avatar and a more or less natural conversation about general issues is conducted. Different emotional colored avatars were used to elicit various emotions of the human interlocutor, for example anger or happiness. A second example is the PIT corpus where two subjects are conducting a dialog to agree on a specific restaurant to go to [56]. A computer assistant with an integrated dialog system is added to the conversation to assist the main user in this process.

The approach for the present and the preceding edition of the AVEC challenge comprises in a sense both approaches as there are different tasks to conduct by the test subjects (for example counting or reading out loudly) but also parts of comparably free speech that is almost like a therapeutic session.

2.4 Conclusions

With this said and acknowledging the scientific progress in the affective computing community, we still argue that the emotional status of a subject is normally hidden in a manifold of different dispositions of the subject and also the circumstances of the distinct recording. This could be due to different display rules or the individual biases of the raters of a sequence. For example the Beck Depression Inventory-II queries many different personal circumstances to determine the depression of a patient that are probably per se not detectable by only watching a video with the subject talking.

Hence, in order to automatically detect the severity of depression of a subject is arguably more feasible to use high-level information about a subject rather than statistical low-level features. Another quite obvious issue that has been already touched earlier is that the procedure of the recording and the degrees of freedom of its annotation strongly determine the a-priori probability of a label and also of the shape of a continuous annotation. In the following these circumstances are used for the development of classification techniques for the continuous affect sub-challenge and the depression sub-challenge.

3. APPROACHES & RESULTS

The classification approaches and the respective results of the numerical evaluations for the two sub-challenges are presented in this section.

3.1 Depression Recognition

For the automatic estimation of the depressive state of the subject, we extracted a variety of high-level features and also a number of coarse features from the raw signals that are provided with the challenge. As classification approach, a Random Forest [1] with 1000 regression trees was used. A Random Forest was chosen because of its robustness against over-fitting and insensitivity against parameter choices.

Concretely, we evaluated the following features for the recognition of the depressive state in a 10-times repeated leave-one-subject-out cross validation (resulting MAE/RMSE values for each feature in parentheses):

1. The id of the subject as a real number as it is provided with the data (10.1130/ 12.6560).
2. The length of the **2.1.** *Freeform* video (9.3160/ 11.1781) **2.2.** *Northwind* video (9.5330/ 11.4917) in frames.
3. The movements of the subject computed by the average pixel difference of two successive images in the
3.1. *Freeform* (9.7410/ 12.3516) and the
3.2. *Northwind* (10.3030/ 12.8922) videos.

4. The variance of the average pixel differences of two images in the **4.1.** *Freeform* (11.1600/ 13.6456) and the **4.2.** *Northwind* (11.448/ 13.6582) videos.
5. The quantiles of the average pixel differences of two images in the **5.1.** *Freeform* (8.5780/ 10.6770) and the **5.2.** *Northwind* (8.2570/ 10.7342) videos.
6. The gender of the subject (10.2410/ 12.2066).
7. Abnormality of the weight of the subject (10.3750/ 12.2651).
8. The estimated age (11.1900/ 13.7971).
9. Estimated information about the socio-economic status of the subject (10.3170/ 12.4397).
10. Ambiance of the recording (10.0910/ 12.0471).
11. Estimated personality trait "facial attractiveness" [29] (10.3140/ 12.3851).
12. Estimated personality trait "likability" [29] (8.9470/ 10.9406).
13. Semantic content of the *Freeform* video (9.8320/ 11.2804).
14. The relative portion of frames for which the Viola-Jones cascade finds a face in the **14.1.** *Freefrom* (10.2920/ 12.5498) and in the **14.2.** *Northwind* (11.0220/ 13.2478) videos.
15. The relative portion of voiced speech frames [3] in the **15.1.** *Freeform* (10.0410/ 11.9037) and in the **15.2.** *Northwind* (9.3460/ 12.2439) videos.
16. Compression ratio of the **16.1.** *Freeform* (9.7150/ 11.9632) and of the **16.2.** *Northwind* (10.6990/ 12.8464) videos with the zip algorithm.
17. Projection on an appearance based high dimensional pixel subspace, whose span constitutes the variations along a manifold that is parametrized by what could be considered a distinct human face computed on the **17.1.** *Freeform* (8.6060/ 10.1949) and the **17.2.** *Northwind* (9.9800/ 11.9970) videos.
18. A posteriori probability of the audio recording of the **18.1.** *Freeform* (10.9530/ 12.8445) and of the **18.2.** *Northwind* (11.3690/ 13.4097) video tested against a hidden Markov model (8 states, 1 Gaussian) that was constructed using publicly available speech.
19. Basic text mining features (i.e., letter appearance statistics) generated via the automatic speech recognition software "Dragon NaturallySpeaking" on the **19.1.** *Freeform* (10.5880/ 12.6451) and on the **19.2.** *Northwind* (9.7930/ 11.5020) videos.
20. The global audio functionals provided with the challenge data for the **20.1.** *Northwind* (9.9540/ 11.8554) and for the **20.2.** *Freeform* (8.4480/ 10.0821) videos.
21. The raw pixel data of the first image of the *Freeform* videos (8.9660/ 10.6994).

An illustration of the connection of the true and the predicted depression scores using features 1–29 in leave-one-speaker-out experiments is shown in Figure 2. It can be seen that there is indeed a roughly linear correlation of the prediction to the label.

Based on this, we compiled 5 different feature bags that were submitted for evaluation on the test partition, which are outlined in Table 1. The feature bags were assembled based on results of a cross validation on the validation set and to investigate the discrimination ability of specific feature groups. The bags were:

1. Features 1, 2.1, 3.1, 14.1, 15.1 and 17.1, which showed the highest scoring results on the validation set.

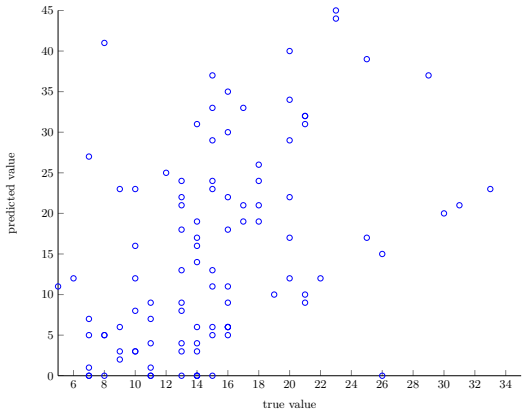


Figure 2: Scatter plot of the predicted value of the depression scores using features $1-29$ against the true value.

2. Meta data only $(1-19)$.
3. Only statistical moments of low-level audio descriptors based on the whole recording (20).
4. Only annotated properties such as gender, approximate age, socio-economic status, background, personality traits and approximate weight class $(6-13)$.
5. 250 best features computed using permuted out-of-bag samples from the Random Forest algorithm.

3.2 Continuous Affective Labels

The estimation of the continuous affective labels has been carried out in a way that follows a different methodology than usual learning tasks. Instead of independently classifying on a per-frame/per-segment basis, task dependent pattern templates were created for each of the three dimensions. The idea is based on the observation, that correct frame/segment decisions are by no means uncorrelated but rather highly dependent on the preceding trajectory and on the fact that the respective performance measure rewards a coarse agreement between the predicted label trajectories and the ground truth more than many individual matching classification results but an overall completely different trajectory (for an illustration of this issue, the reader is referred to Figure 4).

Considering this observation, it is apparent that a change of objectives from estimating closely matching frame-wise values to the robust estimation of the coarse trajectory of the label track is necessary.

The analysis of the label tracks led to a set of distinct prototypical label classes (called proto-labels in the following). We evaluated different approaches to construct feasible proto-labels and found that the following two methods led to models with the best performances on the validation set. The first method is based on support vector regression (SVR) on time-continuous label subsequences. Model parameters were learned to map the time axis to the label values such that the prevailing trajectory is approximated. The use of a suitable kernel function facilitates the regression algorithm to approximate arbitrary shapes of trajectories and in this case leading to very smooth results (compare Figure 3). Adequate use of regularization suppresses oscillations of higher frequency and prevents the deterioration of the learned trajectory. Concretely, an RBF kernel was used

	(1,2.1,3.1,14.1,15.1,17.1)	(1–19)	(20)	(6–13)	250 best oob features	video baseline
RMSE:	9.5802	**9.1880**	9.3488	9.1891	9.7098	10.859
MAE:	7.1400	7.1000	**7.0800**	7.2400	7.2800	8.857

Table 1: Results for the depression sub-challenge on the test partition.

and the respective hyper-parameter was optimized for each task and category using cross-validation on the training set.

The second method is based on an eigenvalue decomposition (EVD) of the trajectory subsequence covariance matrix. The eigenvectors denote different shape variations learned from the label trajectories. The main difference to the SVR method is that the EVD yields an orthogonal decomposition of the trajectory space which can be used to generate proto-labels by superposing the eigenvectors with adequate coefficients. Per task and affective dimension, a cross-validation on the training set is used to determine suitable coefficients for the eigenvectors to maximize correlation between the selected labels and the respective proto-label. The result of the template construction is a fixed curve that is re-sampled to match the length of the respective test video. Examples for concrete templates that are created using SVR and EVD are shown in Figure 3 for the different categories and tasks.

In order to conduct a personalization of the affective recognition templates, a subject clustering approach has been implemented. We used the subset of the features that were developed for the classification of the depressive state comprising features 1 – 29. Based on this, a hierarchical clustering using Ward's distance measure was conducted to divide the available videos into three groups. For each group, both, an EVD and an SVR based template are constructed and for each test sample the more accurate template is chosen based on some hold out data.

In Table 2, the results for the continuous affect recognition sub-challenge are summarized. The 5 submissions are based on the following experimental settings: (1) EVD **and** SVR **with** subject clustering. (2) EVD **without** subject clustering. (3) SVR **without** subject clustering. (4) SVR **with** subject clustering. (5) EVD **with** subject clustering.

4. BOTTOM LINE

This paper outlines our contribution to the AVEC 2014 challenge for the recognition of 3-dimensional affect labels and the depression score for subjects that conduct different tasks in front of a camera. We propose to use background and meta knowledge about the subjects and the respective task that they are going to execute to estimate the annotations.

For the depression sub-challenge a number of additional features were used to gather extra high-level features that could easily be queried from the subject using standard questionnaires (e.g., age, weight, socio-economic status). Another type of feature we used is coarse meta-information that is not directly linked to the depression of a subject at first glance (e.g., the performance of the Viola-Jones cascade on the material or the length of the video material per task). Using this approach we clearly outperformed the baseline results using standard regression techniques.

For the affect recognition sub-challenge we followed a similar approach that reflects the continuous nature in a frame-wise sense of the underlying problem. Hence we developed task and dimension dependent proto-labels that model the course of the video snippets. Two main approaches were proposed for the construction of the labels: one using Eigenvalue decomposition of the training labels and one using SVR for the timely progress of the affective labels, that rendered smoother curves. Additionally a clustering procedure on the information gathered for the depression sub-challenge is conducted in order to partition the subjects into groups based on similarity. For each of these groups, individual proto-labels were constructed and test samples were evaluated using the proto-label that performs better for the respective cluster. Using these approaches the baseline results were clearly outperformed by a factor of almost three.

Please keep in mind that these results were achieved without investigating any kind of low-level features from the audio or video material. Hence the reported errors can be considered as some sort of a-priori curve analogous to using only the a-priori probability in traditional classification approaches. This implies that any approach that uses actual low level/statistical data for the classification has to render higher correlations than the ones outlined in this paper to convincingly show that actual information about the application has been obtained. One reason for this is the choice of the correlation coefficients as performance measure, which is, while well established for the evaluation of statistical regression methods, maybe not the best choice in this application.

5. ACKNOWLEDGMENTS

The authors of this paper are partially funded by 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.

6. REFERENCES

[1] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

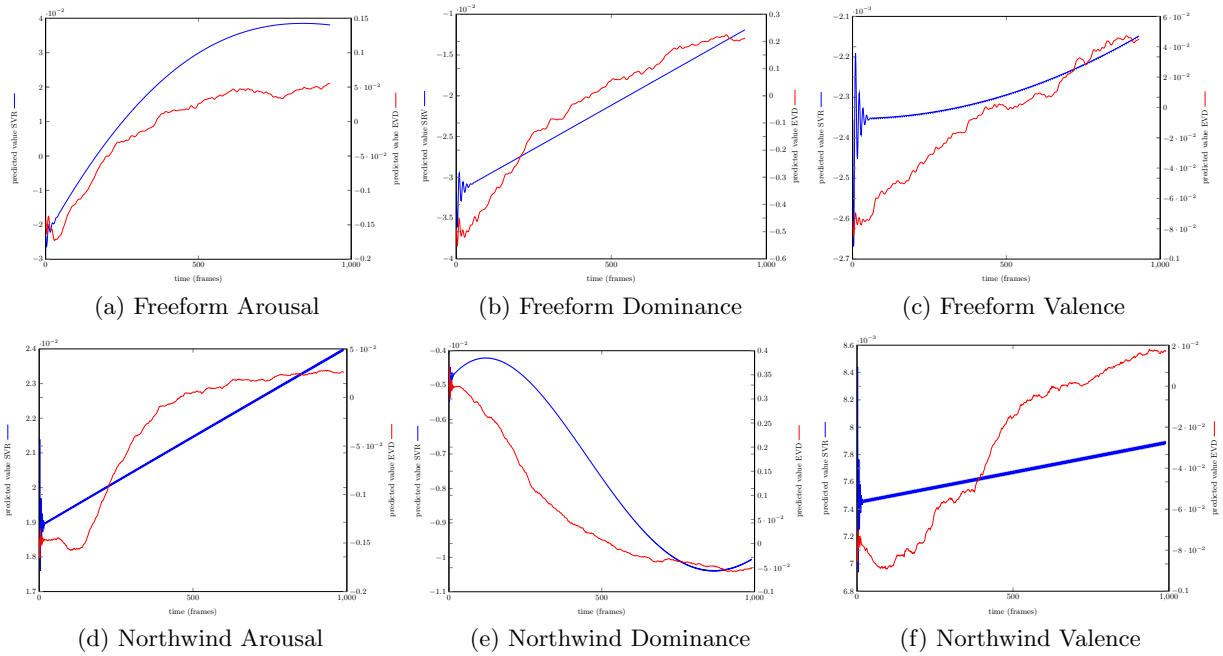
[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Sig. Proc. Mag.*, 18(1):32–80, 2001.

[3] T. Drugman and A. Alwan. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, 2011.

[4] J. Esparza, S. Scherer, and F. Schwenker. Studying self-and active-training methods for multi-feature set emotion recognition. In *Partially Supervised Learning*, pages 19–31. 2012.

[5] M. Glodek, G. Layher, F. Schwenker, and G. Palm. Recognizing human activities using a layered Markov architecture. In *Proc. of ICANN*, volume 7552 of *LNCS*, pages 677–684. Springer, 2012.

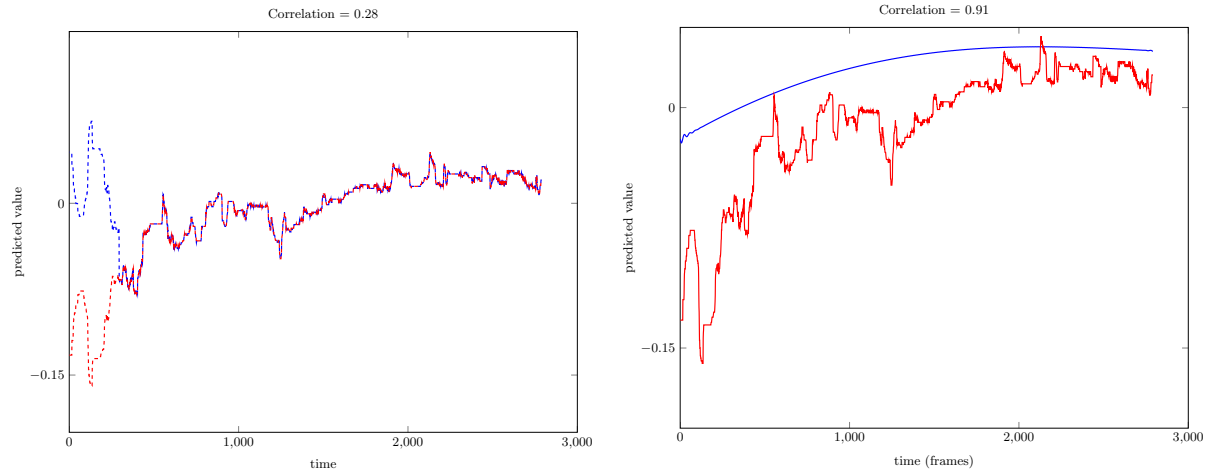
[6] M. Glodek, S. Reuter, M. Schels, K. Dietmayer, and F. Schwenker. Kalman filter based classifier fusion for



(a) Freeform Arousal (b) Freeform Dominance (c) Freeform Valence

(d) Northwind Arousal (e) Northwind Dominance (f) Northwind Valence

Figure 3: Task dependent label templates created with ϵ-SVR (blue) and EVD (red) for the different affective labels.



(a) Strongly overlapping trajectories (b) Coarsely matching trajectories

Figure 4: The left sub-figure illustrates a situation in which the output of a classifier (blue) exactly matches a given label trajectory except for a transient condition in the beginning. Computing the correlation between the two curves yields a value of 0.28 despite the curves being almost identical. The right sub-figure illustrates a contrary example in which the same label is very coarsely estimated by a classifier (again blue). The computed correlation between the curves results in a rather high value of 0.91 although the ground truth curve is touched only occasionally. The coarsely matching trajectory however is enough to yield a high correlation with the label.

	Submission 1	Submission 2	Submission 3	Submission 4	Submission 5	Baseline
Arousal	**0.6330**	0.6013	0.5619	0.6266	0.6229	0.2062
Valence	0.5812	0.5412	0.3048	**0.5869**	0.4609	0.1879
Dominance	**0.5697**	0.5637	0.3931	0.5389	0.5675	0.1959
Mean	**0.5946**	0.5687	0.4199	0.5841	0.5504	0.1966
Mean RMS	0.1009	0.9906	**0.0787**	0.0842	0.1192	n/a

Table 2: Results for the continuous affective sub-challenge on the test partition.

affective state recognition. In *Proc. of MCS*, volume 7872 of *LNCS*, pages 85–94. Springer, 2013.

[7] M. Glodek, M. Schels, G. Palm, and F. Schwenker. Multi-modal fusion based on classification using rejection option and Markov fusion networks. In *Proc. of ICPR*, pages 1084–1087. IEEE, 2012.

[8] M. Glodek, M. Schels, G. Palm, and F. Schwenker. Multiple classifier combination using reject options and markov fusion networks. In *Proc. of ICMI*, pages 465–472. ACM, 2012.

[9] M. Glodek, M. Schels, F. Schwenker, and G. Palm. Combination of sequential class distributions from multiple channels using markov fusion networks. *JMUI*, pages 1–16, 2014.

[10] M. Glodek, S. Scherer, and F. Schwenker. Conditioned hidden Markov model fusion for multimodal classification. In *Interspeech*, pages 2269–2272. ISCA, 2011.

[11] 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 *Proc. of ACII - Part II*, LNCS 6975, pages 359–368. Springer, 2011.

[12] M. F. A. Hady, M. Schels, F. Schwenker, and G. Palm. Semi-supervised facial expressions annotation using co-training with fast probabilistic tri-class svms. In K. I. Diamantaras, W. Duch, and L. S. Iliadis, editors, *Proc. of ICANN*, LNCS 6353, pages 70–75. Springer, 2010.

[13] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[14] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. RASTA-PLP speech analysis technique. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 121–124. IEEE, 1992.

[15] S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, 1998.

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

[17] F. Honold, P. Bercher, F. Richter, F. Nothdurft, T. Geier, R. Barth, T. Hörnle, F. Schüssel, S. Reuter, M. Rau, G. Bertrand, B. Seegebarth, P. Kurzok, B. Schattenberg, W. Minker, M. Weber, and S. Biundo. Companion-technology: Towards user- and situation-adaptive functionality of technical systems. In *Proc. of IE*, pages 378–381. IEEE, 2014.

[18] H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

[19] 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 *Proc. of ICPRAM*, pages 671–678, 2014.

[20] 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 *Proceedings of Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*. Springer, 2014.

[21] M. Kächele and F. Schwenker. Semi-supervised dictionary learning of sparse representations for emotion recognition. In *Partially Supervised Learning*, LNCS, pages 21–35, 2013.

[22] M. Kächele and F. Schwenker. Cascaded fusion of dynamic, spatial, and textural feature sets for person-independent facial emotion recognition. In *Proc. of ICPR*, page (to appear), 2014.

[23] M. Kächele, P. Thiam, G. Palm, and F. Schwenker. Majority-class aware support vector domain oversampling for imbalanced classification problems. In *Artificial Neural Networks in Pattern Recognition*, volume 8774 of *LNCS*, pages 83–92. Springer Berlin Heidelberg, 2014.

[24] 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 *Proc. of ICPR*, page (to appear), 2014.

[25] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.

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

[27] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

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

[29] S. Leikas, M. Verkasalo, and J.-E. Lönnqvist. Posing personality: Is it possible to enact the big five traits in photographs? *Journal of Research in Personality*, 47(1):15 – 21, 2013.

[30] S. Meudt, L. Bigalke, and F. Schwenker. ATLAS – an annotation tool for HCI data utilizing machine learning methods. In *Proc. of APD'12*, pages 5347–5352, 2012.

[31] S. Meudt and F. Schwenker. On instance selection in audio based emotion recognition. In *Proc. ANNPR'12*, pages 186–192, 2012.

[32] S. Meudt, D. Zharkov, M. Kächele, and F. Schwenker. Multi classifier systems and forward backward feature selection algorithms to classify emotional coloured speech. In *Proc. of ICMI*, pages 551–556, 2013.

[33] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. *Neural Computing and Applications*, 9:290–296, 2000.

[34] T. L. Nwe, S. W. Foo, and L. C. De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[35] R. W. Picard. *Affective Computing*. MIT Press Cambridge, 2000.

[36] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice Hall, Eaglewood Cliffs, NJ, 1993.

[37] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proc. of IEEE FG*, pages 1–8, 2013.

[38] D. Rösner, J. Frommer, R. Friesen, M. Haase, J. Lange, and M. Otto. Last minute: a multimodal corpus of speech-based user-companion interactions. In *Proc. LREC'12*, pages 2559–2566, may 2012.

[39] J. A. Russell. A circumplex model of affect. *JPSP*, 39(6):1161–1178, 1980.

[40] M. Schels, M. Glodek, S. Meudt, S. Scherer, M. Schmidt, G. Layher, S. Tschechne, T. Brosch, D. Hrabal, S. Walter, H. Traue, G. Palm, H. Neumann, and F. Schwenker. *Multi-Modal Classifier-Fusion for the Recognition of Emotions*, pages 73–98. Coverbal Synchrony in Human-Machine Interaction. CRC Press, 2013.

[41] 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*, pages 5337–5346, 2012.

[42] 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.

[43] 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. *JMUI*, 8(1):5–16, 2014.

[44] M. Schels, M. Kächele, D. Hrabal, S. Walter, H. C. Traue, and F. Schwenker. Classification of emotional states in a woz scenario exploiting labeled and unlabeled bio-physiological data. In *Proc. of Partially Supervised Learning*, pages 138–147. Springer, 2012.

[45] M. Schels, S. Scherer, M. Glodek, H. Kestler, G. Palm, and F. Schwenker. On the discovery of events in EEG data utilizing information fusion. *Computational Statistics*, 28(1):5–18, 2013.

[46] M. Schels, P. Schillinger, and F. Schwenker. Training of multiple classifier systems utilizing partially labeled sequential data sets. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2011.

[47] K. R. Scherer, T. Johnstone, and G. Klasmeyer. *Handbook of Affective Sciences - Vocal expression of emotion*, chapter 23, pages 433–456. Affective Science. Oxford University Press, 2003.

[48] S. Scherer, M. Glodek, G. Layher, M. Schels, M. Schmidt, T. Brosch, S. Tschechne, F. Schwenker, H. Neumann, and G. Palm. A generic framework for the inference of user states in human computer interaction: How patterns of low level communicational cues support complex affective states. *JMUI*, pages 117–141, 2012.

[49] S. Scherer, M. Glodek, F. Schwenker, N. Campbell, and G. Palm. Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. *ACM Transactions on Interactive Intelligent Systems*, 2(1):1–31, 2012.

[50] M. Schmidt, M. Schels, and F. Schwenker. A hidden markov model based approach for facial expression recognition in image sequences. In *Proc. of ANNPR'10*, LNAI 5998, pages 149–160. Springer, 2010.

[51] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011 — the first international audio visual emotion challenges. In *Proc. of ACII (2011)*, LNCS, pages 415–424, 2011. Part II.

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

[53] F. Schwenker, M. Frey, M. Glodek, M. Kächele, S. Meudt, M. Schels, and M. Schmidt. A new multi-class fuzzy support vector machine algorithm. In *Artificial Neural Networks in Pattern Recognition*, volume 8774 of *LNCS*, pages 153–164. 2014.

[54] F. Schwenker and E. Trentin. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, 37:4–14, 2014.

[55] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3:42–55, April 2012. Issue 1.

[56] P.-M. Strauß, H. Hoffmann, W. Minker, H. Neumann, G. Palm, S. Scherer, F. Schwenker, H. Traue, W. Walter, and U. Weidenbacher. Wizard-of-oz data collection for perception and interaction in multi-user environments. In *Proceedings of LREC*, pages 2014–2017, 2006.

[57] S. Tschechne, G. Layher, and H. Neumann. A Biologically Inspired Model for the Detection of External and Internal Head Motions. In *Proc. of ICANN*, LNCS 8131, pages 232–239, 2013.

[58] 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 *Proc. of ACM Multimedia*. ACM, 2014.

[59] S. Walter, J. Kim, D. Hrabal, S. Crawcour, H. Kessler, and H. Traue. Transsituational individual-specific biopsychological classification of emotions. *IEEE Transactions on Systems, Man, and Cybernetics*, 43(4):988–995, 2013.

[60] S. Walter, S. Scherer, M. Schels, M. Glodek, D. Hrabal, M. Schmidt, R. Böck, K. Limbrecht, H. Traue, and F. Schwenker. Multimodal emotion classification in naturalistic user behavior. In *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, volume 6763 of *LNCS*, pages 603–611. 2011.

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