Chapter 19
Multimodal Affect Recognition in the Context of Human-Computer Interaction for *Companion*-Systems

Friedhelm Schwenker, Ronald Böck, Martin Schels, Sascha Meudt, Ingo Siegert, Michael Glodek, Markus Kächele, Miriam Schmidt-Wack, Patrick Thiam, Andreas Wendemuth, and Gerald Krell

Abstract In general, humans interact with each other using multiple modalities. The main channels are speech, facial expressions, and gesture. But also bio-physiological data such as biopotentials can convey valuable information which can be used to interpret the communication in a dedicated way. A Companion-System can use these modalities to perform an efficient human-computer interaction (HCI). To do so, the multiple sources need to be analyzed and combined in technical systems. However, so far only few studies have been published dealing with the fusion of three or even more such modalities. This chapter addresses the necessary processing steps in the development of a multimodal system applying fusion approaches.

F. Schwenker (✉) • M. Schels • S. Meudt • M. Glodek • M. Kächele • M. Schmidt-Wack • P. Thiam
Institute for Neural Information Processing, University of Ulm, 89069 Ulm, Germany
e-mail: friedhelm.schwenker@uni-ulm.de; martin.schels@uni-ulm.de; sascha.meudt@uni-ulm.de; michael.glodek@uni-ulm.de; markus.kaechele@uni-ulm.de; miriam.schmidt-wack@uni-ulm.de; patrick.thiam@uni-ulm.de

R. Böck • I. Siegert
Cognitive Systems Group, Institute for Information Technology and Communications, Otto von Guericke University, PO Box 4120, 39106 Magdeburg, Germany
e-mail: ronald.boeck@ovgu.de; ingo.siegert@ovgu.de

A. Wendemuth
Cognitive Systems Group, Institute for Information Technology and Communications, Otto von Guericke University, PO Box 4120, 39106 Magdeburg, Germany

Center for Behavioral Brain Sciences, 39118 Magdeburg, Germany
e-mail: andreas.wendemuth@ovgu.de

G. Krell
Technical Computer Science Group, Institute for Information Technology and Communications, Otto von Guericke University, PO Box 4120, 39106 Magdeburg, Germany
e-mail: gerald.krell@ovgu.de

S. Biundo, A. Wendemuth (eds.), *Companion Technology*, Cognitive Technologies,
DOI 10.1007/978-3-319-43665-4_19

ATLAS and ikannotate are presented which are designed for the pre-analyzing of multimodal data streams and the labeling of relevant parts. ATLAS allows us to display raw data, extracted features and even outputs of pre-trained classifier modules. Further, the tool integrates annotation, transcription and an active learning module. Ikannotate can be directly used for transcription and guided step-wise emotional annotation of multimodal data. The tool includes the three mainly used annotation paradigms, namely the basic emotions, the Geneva emotion wheel and the self-assessment manikins (SAMs). Furthermore, annotators using ikannotate can assign an uncertainty to samples.

Classifier architectures need to realize a fusion system in which the multiple modalities are combined. A large number of machine learning approaches were evaluated, such as data, feature, score and decision-level fusion schemes, but also temporal fusion architectures and partially supervised learning.

The proposed methods are evaluated on either multimodal benchmark corpora or on the datasets of the Transregional Collaborative Research Centre SFB/TRR 62, i.e. Last Minute Corpus and the EmoRec Dataset. Furthermore, we present results which were achieved in international challenges.

19.1 Introduction

Successful human-computer interaction (HCI) requires that the computer be able to consider and fulfill different sub-tasks [8] such as perceptual, actuatoric, and cognitive functionalities. In this chapter, we focus on the perceptual sub-system of the computer where pattern recognition methods and machine learning technologies are utilized to perceive the user and to infer the user's affective state [18] which is reflected and represented by perceptible user emotions.

Computers are endowed with various types of sensors to achieve multimodal recognition of affects. These sensors can comprise cameras and microphones to perceive the user's activities in front of the system; laser scanners for localization and tracking of the user; more complex sensors such as eye-tracking devices for detailed gaze analysis; or sensors to observe the user's bio-potentials, e.g. skin conductance (SCR), respiration, electro-cardiogram (ECG), or electroencephalography (EEG). Collecting the raw data in real-time is just the first part of the overall process. In the next step, task-relevant patterns must be identified to enable the system to perform appropriate actions. This process of transferring raw data into a number of classes or categories is known as *pattern recognition*. The idea of pattern recognition is to follow the principle of *learning by example*, utilizing general machine learning algorithms to design classifiers.

Considering the achievements in multimodal disposition recognition we emphasize four main hypotheses being considered in this chapter:

1. The training of multimodal classifiers can benefit from datasets annotated using all available modalities.

2. The training of multimodal classifiers can be conducted in a semi-automatic way without a complete labeling of the training material.
3. Classification performance can be improved by considering the temporal evolution of the observed features.
4. Recognition performance can be improved by applying a multimodal classification approach including a temporal fusion.

Basics in Pattern Recognition In HCI scenarios, technical systems have to combine information from different modalities. This is usually achieved by so-called multiple classifier systems (MCSs) which integrate several classifiers to solve a specific classification problem. The main goal is to obtain a combined output that provides a more accurate and robust classification. In a typical MCS scenario, a complex high-dimensional classification problem is decomposed into smaller subproblems for which improved solutions can be achieved.

In MCS, it is assumed that the raw data X originates from an underlying source, but each classifier receives different subsets of X, e.g. X is applied to multiple types of feature extractors $F_1, \ldots, F_N$ computing multiple views $F_1(X), \ldots, F_N(X)$ of the same raw input data X. Feature vectors $F_j(X)$ are used as the input to the jth classifier, computing an estimate y_j of the class membership of $F_j(X)$. This output y_j might be a crisp class label or a vector of class memberships, e.g. estimates of posterior probabilities. Based on the multiple classifier outputs $y_1, \ldots, y_N$, the combiner produces the final decision y. Combiners can be grouped into fixed transformations of the classifier outputs $y_1, \ldots, y_N$ and trainable mappings. Examples of fixed combining rules are *voting*, *(weighted) averaging* and *multiplying*. By means of an additional optimization procedure, trainable mappings can be realized using the classifier decisions as the inputs to a classifier which performs the final combination. Popular members of this group are artificial neural networks, decision templates and support vector machines [28].

Training a classifier based on vector-valued data can be achieved by computing gradients of error functions with respect to the parameters of some predefined classifier model, such as a multilayer feed forward neural network or a kernel machine. Usually, the raw data comes as a continuous stream of data, e.g. video, audio streams, or waveforms of bio-potentials, and for some tasks the temporal structure of the data might be of importance. Classifier training based on sequential data is much more complex. The structure of the underlying classifier model must be able to process input sequences of different lengths, in particular. In contrast to constant length vectors, sequences may have different lengths even when they represent the same class, e.g. spoken words in speech recognition. Classifier models that often come into play in this context are hidden Markov models (HMMs) and recurrent neural networks (RNNs).

19.2 Machine Learning Framework for Emotion Recognition

A quite important channel of communication between humans is speech, which further allows the transfer of emotional states via content, prosody or paralinguistic cues, which results also in audio-based emotion classification. Another important channel is given by visual perception, which is used by humans to convey their emotional states using facial expressions or body poses. Video-based emotion recognition focuses mainly on the extraction and recognition of emotional information from facial expressions. Attempts have been made to classify emotional states from body and head gestures, as well as from the combination of different visual modalities, such as facial expressions and body gestures which were captured by two separate cameras [16]. Furthermore, emotion recognition can be based on psycho-physiological measurements, e.g. SCR, respiration, ECG, electromyography, or EEG. In contrast to speech, gestures and facial expressions, psycho-physiological measurements are directly generated in the (human) autonomic nervous system and cannot be imitated [24].

The MCS approach is very promising for improving the system's overall classification performance. The individual classifier outputs of the classifier ensemble, which is based on different feature views or modalities, need to be accurate and diverse. While high accuracy is an obvious requirement for members of the ensemble, the concept of classifier diversity is less intuitive to grasp. Members of an ensemble can be regarded as being diverse if the corresponding classifiers disagree on a set of misclassified data [50]. In our work on multimodal emotion recognition, ensemble members have been trained on various types of features extracted mainly from the user's voice and the facial region (e.g. fundamental frequency, Mel-Frequency Cepstral Coefficients (MFCCs), modulation spectrum from the audio signal and form and motion patterns from the video channel) [10, 44]. Besides these more external physical expressions, human emotions (initially studied in human-human interactions) consist of feelings, thoughts and many other types of internal (physiological) processes. Therefore, measuring physiological parameters, such as skin conductivity, heart rate, respiration, or brain activity from EEG, is the first step to studying the automatic recognition of these internal emotional states. The numerical evaluation showed that MCS using fixed and trainable fusion mappings applied to multimodal emotional data can outperform unimodal classifiers. Even in unimodal applications the overall recognition performance increases in many cases by combining outputs of multiple classifiers trained on different features views [39, 50].

Research in facial expression and that in speech-based emotion recognition [35] are usually performed independently from each other. However, in almost all practical applications, people speak and exhibit facial expressions at the same time, and consequently both modalities should be used to perform robust affect recognition. Therefore, multimodal, and in particular audio-visual, emotion recognition has been emerging as a fruitful research topic in recent times [56]. Approaches applying MCS to the classification of human emotions are presented in [6, 42, 45, 49, 58].

19.3 Data Acquisition and Benchmark Data

19.3.1 Survey of Relevant Benchmark Data Sets

In the beginning of affect recognition and, especially emotion or disposition recognition, data sets with acted material were mainly used to create controlled conditions and a setting that allows automated recognition which is not further impeded by difficulties in the extraction process. This results in high quality-data and, further, provides a kind of ground truth in the assigned labels. Such ground truth is important for the validation of recognition systems. For this, the participants of the recording were either actors or naïve speakers who were asked to react in a specific manner. Therefore, a predefined situation is created fixing intended labels by design. In general, it can be assumed that acted material is quite expressive in the shown affects [1]. Such a way of generating corpora was quite similar throughout all modalities (cf. e.g. [59]) since the focus could be set on the development of suitable classifiers and methods. Prominent examples of acted corpora include the Cohn-Kanade facial expression dataset [23], the Berlin database of emotional speech [4] and the Eight-Emotion Sentics Data for biophysiology [17]. To foster the evolution in the various fields, several researchers proposed and conducted challenges. The most prominent are the AVEC challenges—providing audio-visual data sets for various sub-tasks in the emotion recognition, usually concerned with near real-life situations—[56] and the (so-called) EmotiW challenges [7] that emulate a challenging *in-the-wild* setting using emotional snippets from movies.

On the other hand, the question arises: Why shall we consider data sets containing naïve material? For automatic emotion recognition from speech, Batliner et al. [1] discussed the importance of real-life material in detail. This is of interest since it can be assumed that dispositions or affects are expressed in a subtle manner, especially in real-life interactions. Therefore, the research community has to push towards non-acted corpora. In [1] three types are distinguished between based on emotional classes, namely acted, read, and real-life. Besides the characteristic of the data recording, the contained dispositional or affective classes have to be considered (cf. e.g. [55]). The novel shift towards more real-life scenarios is considered in both corpora recorded by the authors and various other research groups.

Prominent naturalistic corpora include the PIT corpus [54], which is conducted as a computer-assisted multi-party dialog, and the RECOLA corpus [38], in which pairs of participants are collaborating to solve a survival task in a Wizard-of-Oz setting. Another notable corpus is the MAHNOB-HCI [53] dataset. The unique part about this data collection is that a multitude of modalities has been recorded: Audio, video, bio-physiology but also EEG and eye gaze, have been recorded.

Besides the mentioned data sets, a strong focus of this work lies on two corpora providing naturalistic interactions with a technical system: the Last Minute Corpus (LMC) (cf. Chap. 13) and the EmoRec corpus [58].

19.3.2 Annotation of Emotional Data

To conduct the emotional labelling, the annotators should be supported by a tool assisting them. Several tools exist to support the literal transcription, for instance EXMARaLDA [46] or Folker [45], but for emotional labelling such tools are rare. For content analysis of videos, the tool Anvil [25] can be used.

19.3.2.1 ATLAS

The freely available ATLAS annotation, labelling and data investigation tool developed at the University of Ulm [31] was designed to assign blocked labels to affective material. In the current version of the ongoing project, it is extended to support fuzzy and fully continuous labelling techniques. Contrary to most other annotation tools, ATLAS is not limited to a maximum number of data streams or specific annotation paradigms. It is possible to depict various recorded raw data like audio, video or, in general, digital sampled values (e.g. bio-physiological signals). Additional extracted features, crisp and probabilistic results of classification procedures and mixed types of labels can be displayed in order to support researchers in obtaining a better understanding of their data, algorithms and results. Synchronous playback of all streams and information is possible.

The presentation complexity of the UI is adaptive to the user's needs. That means that an expert researcher is able to visualize a large amount of detailed information and investigate it at once (see Fig. 19.1), while complex details can also be hidden from unexperienced users to prevent them from being confused. ATLAS is platform-independent and it provides an interface to many common data formats, like MATLAB files. In the case of large datasets a client-server-based distributed annotation structure is implemented, in order to divide computational cost and to give the possibility to annotate with multiple raters at the same time. The annotation supports generic predefined structures that can be tailored to the researchers' needs.

Fig. 19.1 Overview of the adaptive ATLAS UI (expert mode shown). Labels, acoustic properties and physiological raw data are depicted on the *left*. Video and infrared and depth information coupled with their corresponding extracted body data are shown in the *center*. On the *right*, some additional control and detail information windows are presented

Thus ATLAS is not restricted to specific existing emotional models. This leads to the fact that ATLAS is also usable outside the core affective computing community.

Finally, ATLAS includes active learning techniques to provide assistance to the researchers and raters. Externally extracted features can be combined with annotated labels in order to train a classifier. This classifier then suggests additional labels of instances which seem to be most certain, which can be either accepted or rejected by the rater. This information is added to the training information of the next iteration steps. Active learning and semi-supervised learning techniques can improve annotation speed dramatically. As a practicable application of this active learning approach a speaker segmentation tool is available to segment silence and different voices in a quick way.

19.3.2.2 ikannotate

The tool ***i***nterdisciplinary ***k***nowledge-based ***anno***tation ***t***ool for ***a***ided ***t***ranscription of ***e***motions (*ikannotate*) [2] is hosted at the Otto von Guericke University Magdeburg (cf. http://ikannotate.cognitive-systems-magdeburg.de). The particular focus of *ikannotate* is the support of a literal transcription enhanced with phonetic annotations and an emotional labelling using different methods. Both steps are pursued in one tool that provides a more convenient way of data processing. Further, both tools, *ikannotate* and ATLAS, complement each other.

As stated in [2], *ikannotate*'s first release (in 2011) was focused on audio material only. Besides audio processing, the latest version integrates modules which allow a labelling based on visual information, as well. Additionally, several support functions are implemented in *ikannotate*, helping to enrich the labelling or the post-processing. These are, for instance, tagging of dispositionally colored words in the utterance, assigning of uncertainty levels for the labelling, and modules for post-processing like feature extraction.

Enhanced Literal Transcription As it is known from [51], literal transcription is reasonably done on the utterance level since emotions and dispositions change slowly and, thus, one utterance covers one affective state. Therefore, *ikannotate* uses utterances as basic units in transcription. Generally, transcription is done by well-trained experts using tools like Folker [45] or common text editors. Enriching the transcript with prosodic or phonetic annotation usually demands expert knowledge since the utilized annotation paradigms are quite complex. *ikannotate* combines transcription and annotation, and further allows even non-experts to handle audio material properly since the annotator is supported by the tool in both steps.

Necessary information for transcription like start and end times are set by click events. The dialogue structure can be examined by corresponding tabs, and the text input is focused on the current utterance. The annotation is supported by click events as well. Thus, internal complexity of transcription methods in terms of symbols is avoided and annotation is made possible for non-expert users. In *ikannotate*

the Gesprächsanalytisches Transkriptionssystem (dialogue analytic transcription system) is implemented (cf. [2]) including the corresponding features.

Emotional Labelling Another important step in the pre-processing of audio material is the assignment of emotional labels. For this the labeller is assisted by *ikannotate* as well. The tool supports three emotional labelling paradigms: (1) list of emotions (particular emotional phrases are combined in various lists as discussed in Chap. 21), (2) Geneva Wheel of Emotions (GEW) as proposed by Scherer [41], and (3) Self-Assessment Manikins (SAMs) according to [29].

19.4 Context-Aware Temporal Information Fusion Architectures for Multimodal Affect Recognition

Modern fusion architectures for affect recognition have to implement numerous features to take into account the characteristics of emotions. Information about emotions can be gathered from different origins such as multimodality, temporality or the context [15]. Emotions are inherently conveyed by humans using multiple channels [36] which complement each other. The most prominent ones are the auditory and visual channels. Furthermore, emotional analysis can be grouped into categories of different temporal granularities, like expression, attitude, mood, and trait [5]. Thus, affective recognition results have to be temporally combined using a suitable fusion technique. In most cases, and especially in the context of HCI, emotions are related to events or entities in the world [43]. The recognition of this relation is not only crucial to enhance the performance of the classifier system, but it is also of central importance for the further processing in the *Companion*-System. Therefore, affective fusion architectures should incorporate additional user or environmental context.

19.4.1 Fusion of Time-Windowed Features

Audio and video provide a less invasive way to obtain user data for the estimation of affective user states compared to directly user-attached or implanted biometric sensors. However, information on speech, facial expressions, and hand and body gestures [27, 37, 44] is often superimposed by noise and signals unrelated to the affective state to be detected. For instance, facial expression detectors have to cope with problems when a subject turns away, when feature extraction is hampered by wearing glasses, or when mouth movement caused by spoken utterances overlays facial expression. In addition, the target class of the affective user state may be characterized by a vast variety of facial appearances, which makes affective state recognition out of facial expression even more complicated.

Since audio and video are functions over time, we obtain time series of features and intermediate classifier decisions. It is obvious that temporal dependencies between different states exist and that the previous user states have an influence on the current state. We exploited these dependencies by considering the dynamic properties of features [33, 44].

Our investigations showed that linear classifiers often outperform non-linear classifiers [11, 37], since more restrictive classifier functions can be more robust against noise and overfitting in case of training data shortage. Using ensembles of classifiers and exploiting the temporal characteristics of features improved the results significantly. While ensemble learning approaches helped to capture the variety of the target class [3, 12], a time window of features has been applied to consider the dynamics of affective states [37].

The proposed approaches have been examined on the LAST MINUTE corpus (cf. Chap. 13) providing non-acted data of a Wizard-of-Oz setting. In this example, two selected affective user states—namely the normal (Baseline) and the stressed (Challenge) user states—had to be classified by analysing video data of the face. The screenplay of the LAST MINUTE experiment defines the time periods of induced affective state classes to be detected. An extra annotation of ground truth is therefore not required for training data generation because it is directly given by the start and end instants of the events specified in the screenplay. We must be aware that the subject can only be assumed to be in the desired affective state. In reality we try to detect the event according to the screenplay of the experiment and not necessarily the actual affective state.

Figure 19.2 (left) shows the time series of facial measurements creating feature channels for classification. In this case, 13 normalized geometric distances between significant facial points and additionally the eye blink frequency have been collected in a temporal window as input for a linear classifier. The classifier weights for the time series of feature data have been determined in a similar way as for matched

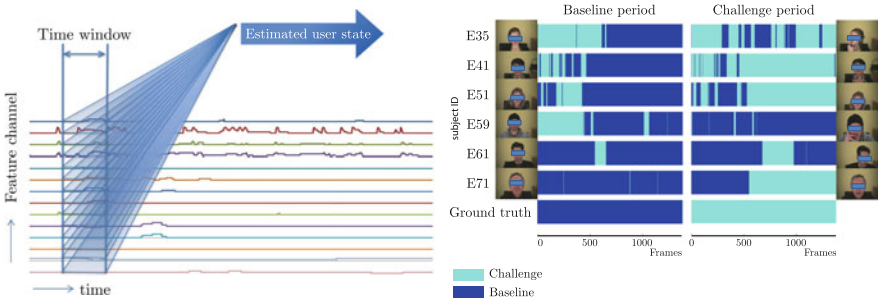

Fig. 19.2 *Left:* Time channels of features are collected in a time window providing the input for user state estimation, in this case 13 geometric features and the eye blink frequency in a time window of 0.6 s. *Right:* Output of linear classification (leave-one-out cross-validation) for six (anonymized) subjects in Baseline (*dark*) and Challenge (*bright*) periods and the ground truth defined by the screenplay of the experiment. Subject ID-related accuracies: E35: 61%, E41: 85%, E51: 77%, E59: 33%, E61: 57%, E71: 80%

filters or deconvolution [26, 34]. Applying a simple threshold to the filter output gives the decision of this linear two-class classifier. The length of the time window, which has been determined for best classification results using leave-one-out cross-validation, comprises 15 frames (0.6 s). A detailed evaluation of the influence of the window size is given in [37].

Figure 19.2 (right) shows the classification results over time for six selected subjects together with the screenplay-defined ground truth values. The recognition accuracy ranges from 33 to 80% for the individual subjects. Depending on such factors as age, biological gender and individual temper, but also on how the subject is used to communicate with a technical system, the facial expressiveness obviously varies a lot by the individual subject. This also holds for the ability of a classifier to detect a user state out of just facial features. Additionally, the unrestricted setting in this scenario does not ensure that the subject is really in the desired affective state (ground truth). Nevertheless, Fig. 19.2 (right) shows that the affective state recognition is capable of distinguishing the given user states to some extent even without creating ensembles of individuals.

An overall classification accuracy of 66% has been calculated in this example, which is quite vague on its own, but may be considered as a typical value for non-acted experimental data. A combination with other modalities is therefore one way to aim for higher confidence in the detected user state.

19.4.2 Temporal Multimodal Fusion Architectures

In the multimodal recognition of affective user states in real-world scenarios, decisions from multiple sources have to be combined. These sources can become inoperative, e.g. due to sensor malfunctions or a missing signal. This issue can be addressed by making use of the temporal process of classifier decisions. Two approaches, namely the Markov fusion network (MFN) and the Kalman filter for classifier fusion, can be applied to perform temporal multimodal fusion.

19.4.2.1 Markov Fusion Networks

The MFN is a probabilistic model for multimodal and temporal fusion introduced in [14]. It is based on a Markov network defined over three potential functions.

A number of $M \times T$ probability distributions over I classes is provided as input to the MFN, where M denotes the number of classifiers generating decisions and T is the number of time steps. The classifiers provide a class distribution for each time step based on each modality. The class probability distribution $m = 1, \ldots, M$ at time step $t \in \mathscr{L}_m$ is a vector $\mathbf{x}_{mt} \in [0, 1]^I$ summing up to 1. Since a classifier decision might be temporally missing, the set $\mathscr{L}_m$ contains only the time steps in which class probability distributions of the classifier m are available. Assuming, without loss of generality, that the probability distributions are available for all time steps, the

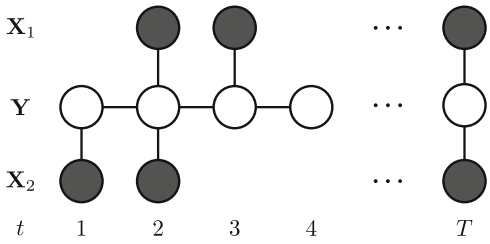

Fig. 19.3 Graphical model of the MFN. The sequence of combined estimates $\mathbf{y}_t$ is influenced by the available decisions $\mathbf{X}_{mt}$ of the source m and $t \in \mathscr{L}_m$ and adjacent combined estimates $\mathbf{y}_{t-1}$ and $\mathbf{y}_{t+1}$

MFN integrates the classifier predictions $\mathbf{X}_m \in [0,1]^{I \times T}$ to a combined estimate $\mathbf{Y} \in [0,1]^{I \times T}$ by making use of two main objectives. The first objective states that the combined estimated probability distributions will be similar to the provided class probability distributions. The second objective states that the estimated probability distributions are similar in the temporal proximity.

At first, the MFN defining function enforces the estimates to be similar to the observed class probability distributions. The second objective is implemented by temporally connecting the estimates in a Markov chain. The MFN reconstructs regions without classifier decisions by propagating information along the Markov chain.

Figure 19.3 depicts the graphical model of an exemplary MFN, which integrates two sequences of classifier decisions $\mathbf{X}_1$ and $\mathbf{X}_2$ to a combined estimate $\mathbf{Y}$. The classifier decisions are connected to the corresponding estimates in each time step. Whenever a class distribution is unavailable, the input node and the connecting link are omitted. The estimates themselves are temporally connected by the Markov chain.

19.4.2.2 Kalman Filter Architectures

The Kalman filter for classifier fusion operates on the same input data as the MFN [11]. However, the studied implementation was restricted to a two-class classification problem.

The conventional Kalman filter [22] is a well-known algorithm to enhance the quality of noisy measurements over time. It is commonly applied in the field of navigation and object tracking. Instead of calculating a rather simple average of measurements, the Kalman filter explicitly models the measurement noise. The modeled uncertainty can significantly enhance the quality of tracking. The Kalman filter itself is closely related to the HMM, but uses a Markov chain of continuous latent variables.

In [11] the Kalman filter was first applied to classifier decision fusion over time and was extended in order to handle missing classifier decisions. The Kalman filter for classifier fusion approach was tested on the AVEC 2013 challenge data [20] which is discussed in Sect. 19.5.1.

19.4.3 Integration of Context Information

The integration of context information is challenging for machine learning [15]. For this, the hierarchical structure of temporal patterns with different complexities is exploited. For instance, social signals can be decomposed into short-term behavioral cues [43, 57]. However, since no dataset has been recorded so far with a hierarchically structured ground truth of affective states, basic research on the integration of context information was conducted in the field of activity recognition. Several approaches have been proposed, like conditioned hidden Markov model (CHMM), unidirectional layered architecture (ULA), and HMM/CHMM using graph probability densities (HMM/CHMM-GPD).

The CHMM is an extension of the classic HMM, in which the hidden states are influenced by an additional sequence of causes. These causes can be provided by an external classifier decision which serves as additional context information [13].

CHMM can be studied as part of the ULA [13], in which each layer recognizes different classes with increasing complexities. The lowermost layer operates on features derived from the sensors and recognizes basic entities. The subsequent layers operate on the output of the preceding layers, which is given by their probabilistic class predictions. User preferences were recognized in the uppermost layer using a dynamic Markov logic network (DMLN) which models context information in the form of probabilistic logical rules. Studies on the baseline (cf. Sect. 19.4.1) investigate the propagation of context information down to the lowermost layer in order to influence the CHMM.

19.5 Multimodal Affect Recognition Results

19.5.1 Public Benchmark

AVEC Results The two sub-challenges of the 2013/2014 edition of the AVEC challenge comprise the two-/three-dimensional continuous affect sub-challenge and the discrete depression sub-challenge. The dataset contains 150 audio-visual recordings of participants of a clinical study in an inquiry-response cycle in front of a consumer notebook. The task was to estimate the continuous label traces for a test set of 50 videos. Performance was measured using (the magnitude of) Pearson's correlation coefficient, averaged over the test videos. For the 2013 edition of the challenge, a recognition system was developed that combines the input of a hierarchical classifier (consisting of multiple individual SVR and MLP classifiers) for the video modality with a diverse set of audio features using a Kalman filter (see Sect. 19.4.2.2) [20]. The results can be seen in Fig. 19.4 (left). For the 2014 edition, a slightly different approach has been taken. Based on the annotated trajectories, prototypical label traces were created using PCA and SVR to highlight difficulties in the annotation process (i.e. arbitrary starting point and subsequent transient phases

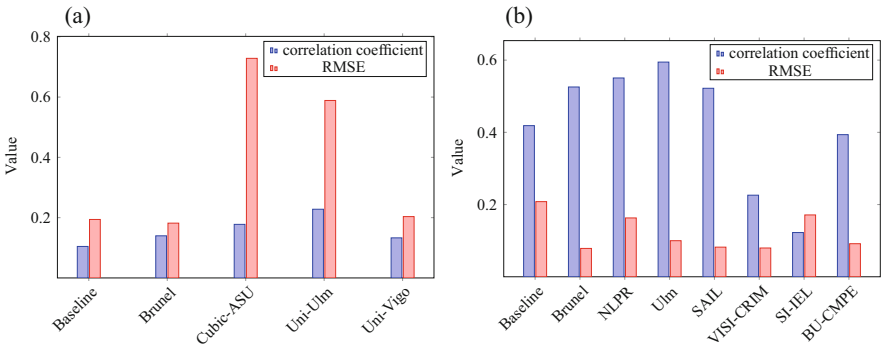

Fig. 19.4 Average correlation coefficient and RMSE of the affect sub-challenge of the 2013 and 2014 editions of the AVEC challenges. Source: http://sspnet.eu/avec2013/ and http://sspnet.eu/avec2014/. (**a**) Challenge results of AVEC 2013. (**b**) Challenge results of AVEC 2014

in the beginning) and of the performance measures. Combined with a clustering to reveal participant groups, the approach led to superior results and the win of the affect sub-challenge. For details the reader is referred to Fig. 19.4 and [21].

EmotiW Results The Emotion Recognition in the Wild challenge [7] focuses on audio-visual emotion recognition from movie snippets extracted from feature films. The snippets offer unconstrained movement, difficult lighting and speech overlapped with music and background noise. To tackle the challenge, the authors presented an approach that combined basic audio features with feature selection and achieved, using only a single modality, competitive results in the 2013 edition of the challenge [32]. Building on this success, in the follow-up challenge the application of enhanced auto-correlation features for the recognition of emotions from speech was proposed [30].

19.5.2 Results

Last Minute Corpus The LMC material has the advantage that a fusion of multimodal classifications can be combined in the context of more subtle dispositions like "concentration" and "thinking". For this, novel characteristics such as self-touching and eye-blink frequency can be observed. These approaches are pursued based on the main idea presented in Fig. 19.5 focussing on two situations, namely baseline (BSL) and weight limit barrier (WLB) [9].

In [52] these novel characteristics are investigated with a focus on the analysis of facial expressions. For the generation of a visual-based classifier the following features are considered: mouth deformations, eye-blink, eyebrow movement, and the general movement of the head (global) as the most prominent and reliably detectable features for the face as used in [27, 52]. On the one hand, common

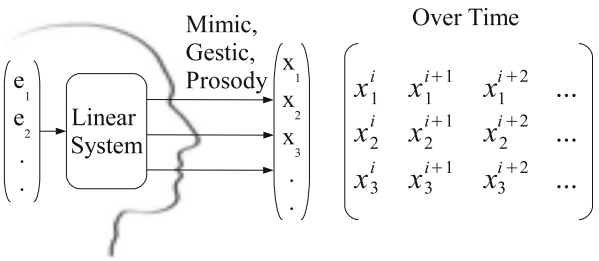

Fig. 19.5 The user, modeled as a linear system, transmits emotional signals via various modalities over several time steps. These inputs can be collected and combined for further processing. The figure is adapted from [37]

features were used to detect mouth movements and eye-blinks [52]. Further, extracted visual features of hand gestures are important, from a psychological point of view, especially "self-touch" and "no-self-touch" when the subject touches his face. Therefore, using skin color and connected component analysis, the overlap of hand and face regions can be detected. For analysing purposes, we investigated 13 subjects of the LMC for visual classification due to different illumination, head positions, and occlusions. According to [37, 52] the processing is as follows: A time window with the size of ten frames (0.4 s) is considered instead of a decision based on individual frames, as it is assumed that the bodily response, which is reflected in changes of the features, is short-time stable.

The visual classification is further combined with an acoustic evaluation. Therefore, an automatic classification system was trained on the material of the same 13 subjects. For evaluation, we applied a leave-one-subject-out (LOSO) strategy and used HMMs/GMMs with common MFCCs with delta and acceleration. This results in an overall mean of the weighted accuracy of 76.01% (std. dev. 6.45%) for the two-class problem.

Based on these classifications a fusion was conducted. For this, an MFN with the following parameters was used: W = 1000, kf = 0.5, kp = 4, kg = 4. The uni-modal classifiers are the facial expressions, gestural analysis, and the acoustic classifier. As pointed out in [52], each modality possesses its own distinct characteristic distribution of decisions over time. The recognition of the emotional state based on facial expressions requires the subject's face to be in the view of the camera. However, in case the subject turns away, a decision may become infeasible. A similar problem occurs in prosodic analysis since it can be performed only if the subject produces an utterance. In the given setting, the decisions derived from the gestural analysis are even more demanding, because they only give evidence for the class WLB. The classifier based on facial expression provides decision probabilities for all frames, the acoustic analysis only for 15.9% of the frames and a gestural analysis only for 9% of the frames. The overall average accuracy is 85.29% (std. dev. 14.22%).

EmoRec The experimental validation can be divided into uni-modal and multi-modal approaches since audio, video, and bio-physiology of the user were recorded. Grounded on the experimental design, a small set of representative experimental sequences (ESs) is selected for classification. ES 2—the experimental part linked to

"positive pleasure, low arousal, high dominance" feeling of the user—and ES 5—referring to "negative pleasure, high arousal, low dominance"—are selected based on their location as opposing octants in the VAD space. For a more detailed view of the experimental setting, we refer you to [58].

The video modality is pre-processed as follows: First, face detection is done based on Viola and Jones' boosted Haar cascade. Salient facial points are detected using a constraint local model followed by an alignment procedure based on selected points. The selected features are: optical flow, motion histograms, pyramids of histograms of oriented gradients and local binary patterns.

Using the diverse information captured by the different feature sets, different fusion methods are employed. First, an ensemble of Support Vector Machines (SVMs) with softmax output was trained on bootstrapped subsets of the training data for each of the four feature sets. The results of each ensemble were aggregated by a trainable combiner in the form of a multilayer perceptron. To compensate for different time resolutions of the features, a common reference time window of 2 s is used to integrate the per-channel decisions. Finally, in another trainable fusion mapping the estimates of the individual channels are aggregated into the final decision. Using this fusion scheme, a final accuracy of 69.2% can be achieved. For an overview of the results (including other fusion methods), the reader is referred to Fig. 19.6. More details are given in [19] as these results were obtained only on a subset of 11 people.

Further results are obtained by analyzing the bio-physiological channels. Since the recorded modalities are inherently different, each channel has to be individually preprocessed. For example, to extract information from the blood volume pulse, first the heartbeats in the form of so-called QRS complexes have to be located. After

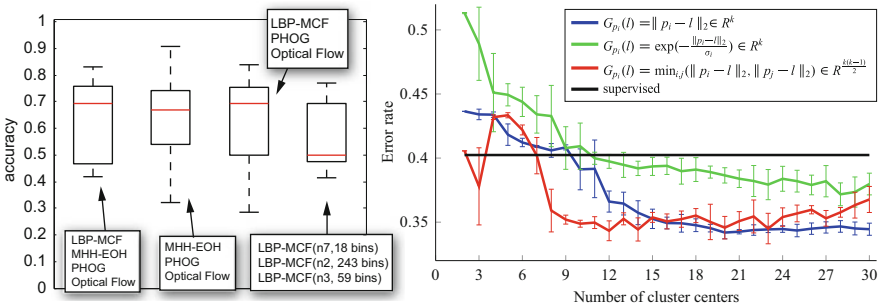

Fig. 19.6 *Left*: The final results obtained with the fusion architecture with different settings. The highest median classification rate was 69.2%, achieved with a combination of all channels. Other channel selections also performed relatively well. Fusing results of the same channel with different settings (bin and neighbourhood size) did not result in an improvement. *Right*: Classification error with standard deviation computed over ten runs. Different cluster techniques are used to augment the classification process. The supervised reference is given as a *black line*. The unsupervised techniques outperform the supervised approach given enough model complexity (here: number of cluster centers)

additional filtering and detrending, different features are computed for each channel on suitable time windows. For the heart rate, well-known statistical features such as the standard deviation, RMSSD, or pNN50 are applied. Additionally, the non-linear features that approximate entropy, recurrence rate and dimensions of the ellipse in a Poincaré plot are calculated. Finally, three features based on the power spectrum density are computed.

The experimental results based on the individual channels suggest that robust recognition is only possible to a certain extent, i.e. the recognition rate is only slightly above chance level (for details, the reader is referred to [40]). To alleviate this problem, the following means are introduced: A fusion step is introduced to combine the individual bio-physiological channels and, additionally, a technique from the field of semi-supervised learning is used to incorporate the additional data provided by the remaining experimental sequences (i.e. ES 1,3,4,6). In the procedure, an unsupervised pre-processing step is used as transformation into a data-driven representation. In Fig. 19.6, the results are summarized. The fusion itself improves the result to an error rate of about 40%. In addition, the unsupervised pre-processing step further lowers the error rate to about 35%.

19.5.3 Active Learning in an HCI Scenario

Plausible annotation of multimodal HCI data is an enormous problem due to the fact of the time-consuming and sensitive annotation process. Furthermore, emotional reactions are often very sparse, resulting in a large annotation overhead to gather the interesting moments of a recording. Active learning techniques provide methods to improve the annotation processes since the annotator is asked to label only the relevant instances of a given dataset.

The approach of active learning was applied on an interaction data set, described in Chap. 12 and published in [48]. A number of subjects were recorded while performing a search task on a screen. They could interact with the system via speech or touch commands. During the search period, the subjects tended to react just a little or not at all emotionally. Usually, all expressive reactions occurred when the subject failed to solve the task. Table 19.1 shows the imbalance of feature instances of the neutral and emotional behaviors.

Table 19.1 Number of neutral and emotional feature instances, enlisted for each participant

ID	12	15	17	23	26	30	Sum.
#Neutral	3387	3365	4149	3406	3409	4489	22,205
#Emotion	108	135	66	208	429	665	1661
Responsiveness	Good	Good	Moderate	Good	Moderate	Moderate	

The last column shows the estimation of the participant's emotional expressiveness

For feature extraction, the facial region was located and extracted within each single frame of a video sequence. Subsequently, each face was divided into a fixed number of non-overlapping blocks. The *Local Binary Pattern* operator on *Three Orthogonal Planes (LBP-TOP)* [60] was applied to each cuboid consisting of each block of facial region for an entire sequence to generate the corresponding histogram of descriptors. These histograms were concatenated to form the feature vector for each sequence.

The feature vectors were utilized as input for a *One Class Support Vector Machine (OCSVM)* [47]. It is an extension of the binary SVM, defining a decision function that takes the value $+1$ in a small region capturing most of the instances, and -1 elsewhere. The instances of the target class are mapped into a Hilbert space H and subsequently separated from the origin by a hyperplane with maximum margin. Consequently, data objects are classified either as outliers or as belonging to the target class. The target class was designated as neutral, because of the extreme imbalance in the data; therefore the outliers are the sparse emotional moments.

The OCSVM was used for an active learning approach. The initial adjustment of the classifier pooled all data points. The first step presented the data points, which were explained least by the model (depending on their distance to the hyperplane), to an expert to be labeled. The labeling information was utilized to improve the classifier's performance. Both steps were applied in a loop several times.

In each iteration, the ten worst data points (outliers) were presented to an annotator. In the first iteration steps, most of the outliers represented emotional moments and were labeled accordingly. During the further course, the number of data points differing from neutral decreased. After 50 iterations (500 data points) the process was stopped. Figure 19.7 shows this behavior for six participants.

Moreover, a closer look shows that 82.41% of the emotional moments of participant 12 were identified by labeling just 14% of the entire dataset. The same observations can be made for participant 15: 74.07% of the outliers were identified by labeling about 14.3% of the entire dataset.

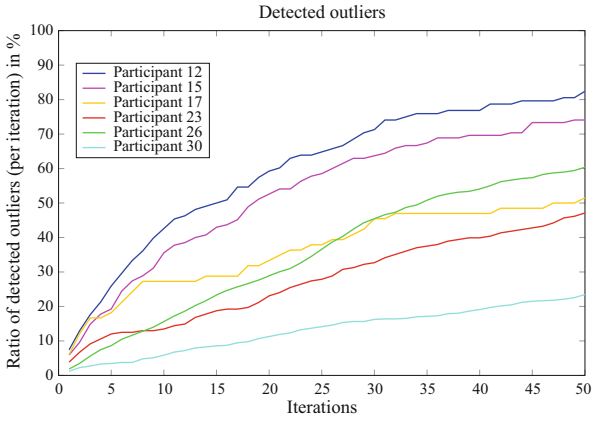

Fig. 19.7 Number of data points labeled as emotional moments in relation to all emotional moments in the data set for each iteration step. As can be seen, about 82.41% of emotional events could be detected for participant 12 by labeling just 14% of the data

In a subsequent experiment a binary SVM was trained on the 500 labeled data points and achieved a g-mean value[1] of about 0.8. To compare this performance with common methods, the whole data set was manually labeled and classified with a binary SVM, trained on all data points. The accomplished g-mean was almost the same. The reason for this is that most of the support vectors lie within the labeled 500 data points and the remaining points do not contribute to the decision hyperplane. Hence, the proposed active learning approach generates an effective classification model while labeling just a small portion of the dataset.

19.6 Conclusion

The multimodal recognition of user affects is a challenging issue which could be faced by the combination of various modalities. Besides the combination of suitable modalities and features, effort should be applied to a stringent development and handling of recognition architectures. In this chapter, we presented an overview of our work done in the context of affect recognition based on multiple sources. Especially, the use of active learning approaches and Markov fusion networks provides an improvement in the processing of naturalistic, affectively afflicted material. The achieved results are either based on corpora recorded in the authors' research groups or on publicly available benchmark data sets. For both categories remarkable results were obtained, in particular in the benchmark challenges (cf. Sect. 19.5.1). Based on such achievements, we processed the internally recorded data sets—namely LMC (cf. Chap. 13) and EmoRec [58]—showing that the discussed recognizer architectures can be applied in naturalistic scenarios. In particular, the use of temporal and contextual information for multimodal fusion improved the classification (cf. Hypotheses 2 and 4).

Besides the classification issue, the preprocessing of data should also be approached multimodally (cf. Hypothesis 3). Therefore, we introduced two annotation tools assisting during the annotation and labelling. *ikannotate* mainly focuses on the annotation and labelling based on audio and video streams. In addition, ATLAS allows a synchronous handling of audio, video, and biophysiological data. Furthermore, it applies active learning techniques to assist in the labelling process. From the active learning perspective, we can conclude that the multimodal classifiers can be established without completely annotated material. The starting point is a small subset which provides reasonable class information. Iteratively, a classifier can be trained while simultaneously annotating the material (cf. Hypothesis 1). Finally, we briefly discussed that data which is recorded synchronously can improve the performance of the classification in terms of multimodal investigations. As is elaborated on in Chap. 22, there are several

[1]The g-mean was chosen because of the strong imbalance between the two classes.

ways of establishing synchronous recordings. For real-world applications, we refer to Chap. 22.

Acknowledgements We thank our highly regarded deceased colleague and friend Prof. Dr. Bernd Michaelis who contributed to the SFB on various topics and provided well-informed suggestions. This work was done within the Transregional Collaborative Research Centre SFB/TRR 62 "*Companion*-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Batliner, A., Fischer, K., Huber, R., Spiker, J., Nöth, E.: Desperately seeking emotions: Actors, wizards and human beings. In: Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, pp. 195–200 (2000)
2. Böck, R., Siegert, I., Haase, M., Lange, J., Wendemuth, A.: ikannotate - a tool for labelling, transcription, and annotation of emotionally coloured speech. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.C. (eds.) Proceedings of ACII. Lecture Notes on Computer Science, vol. 6974, pp. 25–34. Springer, Berlin (2011)
3. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
4. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Proceedings of Interspeech 2005, pp. 1517–1520 (2005)
5. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. **18**(1), 32–80 (2001)
6. Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Netw. **18**(4), 407–422 (2005)
7. Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion recognition in the wild challenge 2014: baseline, data and protocol. In: Proceedings of ICMI, pp. 461–466. ACM, New York (2014)
8. Dix, A., Finlay, J., Abowd, G., Beale, R.: Human-computer Interaction. Prentice-Hall, Upper Saddle River, NJ (1997)
9. Frommer, J., Michaelis, B., Rösner, D., Wendemuth, A., Friesen, R., Haase, M., Kunze, M., Andrich, R., Lange, J., Panning, A., Siegert, I.: Towards emotion and affect detection in the multimodal last minute corpus. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of LREC. ELRA, Paris (2012)
10. Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., Schwenker, F.: Multiple classifier systems for the classification of audio-visual emotional states. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.C. (eds.) Proceedings of ACII - Part II, Lecture Notes on Computer Science, vol. 6975, pp. 359–368. Springer, Berlin (2011)
11. Glodek, M., Reuter, S., Schels, M., Dietmayer, K., Schwenker, F.: Kalman filter based classifier fusion for affective state recognition. In: Zhou, Z.H., Roli, F., Kittler, J. (eds.) Multiple Classifier Systems (MCS). Lecture Notes on Computer Science, vol. 7872, pp. 85–94. Springer, Berlin (2013)
12. Glodek, M., Schels, M., Schwenker, F.: Ensemble Gaussian mixture models for probability density estimation. Comput. Stat. **27**(1), 127–138 (2013)
13. Glodek, M., Geier, T., Biundo, S., Palm, G.: A layered architecture for probabilistic complex pattern recognition to detect user preferences. J. Biol. Inspired Cognitive Archit. **9**, 46–56 (2014)

14. Glodek, M., Schels, M., Schwenker, F., Palm, G.: Combination of sequential class distributions from multiple channels using Markov fusion networks. J. Multimodal User Interfaces **8**(3), 257–272 (2014)
15. Glodek, M., Honold, F., Geier, T., Krell, G., Nothdurft, F., Reuter, S., Schüssel, F., Hörnle, T., Dietmayer, K., Minker, W., Biundo, S., Weber, M., Palm, G., Schwenker, F.: Fusion paradigms in cognitive technical systems for human-computer interaction. Neurocomputing **161**, 17–37 (2015)
16. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. J. Netw. Comput. Appl. **30**(4), 1334–1345 (2007)
17. Healey, J.: Wearable and automotive systems for affect recognition from physiology. Ph.D. thesis, MIT (2000)
18. Hudlicka, E.: To feel or not to feel: The role of affect in human-computer interaction. Int. J. Hum.-Comput. Stud. **59**(1-2), 1–32 (2003)
19. Kächele, M., Schwenker, F.: Cascaded fusion of dynamic, spatial, and textural feature sets for person-independent facial emotion recognition. In: Proceedings of ICPR, pp. 4660–4665 (2014)
20. Kächele, M., Glodek, M., Zharkov, D., Meudt, S., Schwenker, F.: Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In: De Marsico, M., Tabbone, A., Fred, A. (eds.) Proceedings of ICPRAM, pp. 671–678. SciTePress, Setúbal (2014)
21. Kächele, M., Schels, M., Schwenker, F.: Inferring depression and affect from application dependent meta knowledge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, pp. 41–48. ACM, New York (2014)
22. Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Fluids Eng. **82**(1), 35–45 (1960)
23. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Automatic Face and Gesture Recognition, 2000, pp. 46–53 (2000)
24. Kim, K., Bang, S., Kim, S.: Emotion recognition system using short-term monitoring of physiological signals. Med. Biol. Eng. Comput. **42**(3), 419–427 (2004)
25. Kipp, M.: Anvil - a generic annotation tool for multimodal dialogue. In: INTERSPEECH-2001, Aalborg, Denmark, pp. 1367–1370 (2001)
26. Krell, G., Niese, R., Al-Hamadi, A., Michaelis, B.: Suppression of uncertainties at emotional transitions — facial mimics recognition in video with 3-D model. In: Richard, P., Braz, J. (eds.) Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 537–542 (2010)
27. Krell, G., Glodek, M., Panning, A., Siegert, I., Michaelis, B., Wendemuth, A., Schwenker, F.: Fusion of fragmentary classifier decisions for affective state recognition. In: MPRSS, Lecture Notes on Artificial Intelligence, vol. 7742, pp. 116–130. Springer, Berlin (2012)
28. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, New York (2004)
29. Lang, P.J.: Behavioral Treatment and Bio-Behavioral Assessment: Computer Applications, pp. 119–137. Ablex Publishing, New York (1980)
30. Meudt, S., Schwenker, F.: Enhanced autocorrelation in real world emotion recognition. In: Proceedings of the 16th International Conference on Multimodal Interaction, ICMI '14, pp. 502–507. ACM, New York (2014)
31. Meudt, S., Bigalke, L., Schwenker, F.: Atlas – an annotation tool for HCI data utilizing machine learning methods. In: International Conference on Affective and Pleasurable Design (APD'12), pp. 5347–5352 (2012)
32. Meudt, S., Zharkov, D., Kächele, M., Schwenker, F.: Multi classifier systems and forward backward feature selection algorithms to classify emotional coloured speech. In: International Conference on Multimodal Interaction, ICMI 2013, pp. 551–556. ACM, New York (2013)

33. Niese, R., Al-Hamadi, A., Heuer, M., Michaelis, B., Matuszewski, B.: Machine vision based recognition of emotions using the circumplex model of affect. In: Proceedings of the International Conference on Multimedia Technology (ICMT), pp. 6424–6427. IEEE, New York (2011)
34. North, D.O.: An analysis of the factors which determine signal/noise discrimination in pulsed-carrier systems. Proc. IEEE **51**(7), 1016–1027 (1963)
35. Oudeyer, P.: The production and recognition of emotions in speech: features and algorithms. Int. J. Hum.-Comput. Stud. **59**(1-2), 157–183 (2003)
36. Palm, G., Glodek, M.: Towards emotion recognition in human computer interaction. In: Esposito, A., Squartini, S., Palm, G. (eds.) Neural Nets and Surroundings, vol. 19, pp. 323–336. Springer, Berlin (2013)
37. Panning, A., Siegert, I., Al-Hamadi, A., Wendemuth, A., Rösner, D., Frommer, J., Krell, G., Michaelis, B.: Multimodal affect recognition in spontaneous HCI environment. In: 2012 IEEE International Conference on Signal Processing, Communication and Computing, pp. 430–435. IEEE, New York (2012)
38. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8 (2013)
39. Schels, M., Scherer, S., Glodek, M., Kestler, H., Palm, G., Schwenker, F.: On the discovery of events in EEG data utilizing information fusion. Comput. Stat. **28**(1), 5–18 (2013)
40. Schels, M., Kächele, M., Glodek, M., Hrabal, D., Walter, S., Schwenker, F.: Using unlabeled data to improve classification of emotional states in human computer interaction. J. Multimodal User Interfaces **8**(1), 5–16 (2014)
41. Scherer, K.R.: What are emotions? and how can they be measured? Soc. Sci. Inf. **44**, 695–729 (2005)
42. Scherer, S., Schwenker, F., Palm, G.: Classifier fusion for emotion recognition from speech. In: Advanced Intelligent Environments, pp. 95–117. Springer, Boston (2009)
43. Scherer, S., Glodek, M., Layher, G., Schels, M., Schmidt, M., Brosch, T., Tschechne, S., Schwenker, F., Neumann, H., Palm, G.: A generic framework for the inference of user states in human computer interaction: how patterns of low level behavioral cues support complex user states in HCI. J. Multimodal User Interfaces **6**(3–4), 117–141 (2012)
44. Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G.: Spotting laughter in natural multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data. ACM Trans. Interactive Intell. Syst. **2**(1), 4:1–4:31 (2012)
45. Schmidt, T., Schütte, W.: FOLKER: an annotation tool for efficient transcription of natural, multi-party interaction. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (2010)
46. Schmidt, T., Wörner, K.: EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics **19**, 565–582 (2009)
47. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support vector method for novelty detection. In: NIPS, vol. 12, pp. 582–588 (1999)
48. Schüssel, F., Honold, F., Weber, M., Schmidt, M., Bubalo, N., Huckauf, A.: Multimodal interaction history and its use in error detection and recovery. In: Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI'14), pp. 164–171. ACM, New York (2014)
49. Schwenker, F., Scherer, S., Magdi, Y.M., Palm, G.: The GMM-SVM supervector approach for the recognition of the emotional status from speech. In: ICANN (1), Lecture Notes on Computer Science, vol. 5768, pp. 894–903. Springer, Berlin (2009)
50. Schwenker, F., Scherer, S., Schmidt, M., Schels, M., Glodek, M.: Multiple classifier systems for the recognition of human emotions. In: Multiple Classifier Systems, Lecture Notes on Computer Science, vol. 5997, pp. 315–324. Springer, Berlin (2010)

51. Sezgin, M.C., Gunsel, B., Kurt, G.: Perceptual audio features for emotion detection. EURASIP J. Audio Speech Music Process. **2012**, 1–21 (2012)
52. Siegert, I., Glodek, M., Krell, G.: Using speaker group dependent modelling to improve fusion of fragmentary classifier decisions. In: Proceedings of the International IEEE Conference on Cybernetics (CYBCONF), pp. 132–137. IEEE, New York (2013)
53. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **3**, 42–55 (2012).
54. Strauß, P.M., Hoffmann, H., Minker, W., Neumann, H., Palm, G., Scherer, S., Schwenker, F., Traue, H., Walter, W., Weidenbacher, U.: Wizard-of-oz data collection for perception and interaction in multi-user environments. In: Proceedings of LREC, pp. 2014–2017 (2006)
55. Traue, H.C., Ohl, F., Brechmann, A., Schwenker, F., Kessler, H., Limbrecht, K., Hoffman, H., Scherer, S., Kotzyba, M., Scheck, A., Walter, S.: A framework for emotions and dispositions in man-companion interaction. In: Rojc, M., Campbell, N. (eds.) Converbal Synchrony in Human-Machine Interaction, pp. 98–140. CRC Press, Boca Raton (2013)
56. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M.: AVEC 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of ACM MM, AVEC '14, pp. 3–10. ACM, New York (2014)
57. Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social signal processing: state-of-the-art and future perspectives of an emerging domain. In: Proceedings of the International ACM Conference on Multimedia (MM), pp. 1061–1070. ACM, New York, NY (2008)
58. Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H.C., Schwenker, F.: Multimodal emotion classification in naturalistic user behavior. In: Jacko, J.A. (ed.) Proceedings of the 14th International Conference on Human Computer Interaction (HCI'11), Lecture Notes on Computer Science, vol. 6763, pp. 603–611. Springer, Berlin (2011)
59. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
60. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)