From Synchronous to Asynchronous Event-driven Fusion Approaches in Multi-modal Affect Recognition

DISSERTATION

zur Erlangung des akademischen Grades
Doktor der Informatik

Lehrstuhl für Multimodale Mensch-Technik Interaktion
Universität Augsburg

M.Sc. Florian Lingenfelser

2018

Datum der Disputation	Augsburg, 22.02.2018
Erstgutachter	Prof. Dr. Elisabeth André
Zweitgutachter	Prof. Dr. Björn Schuller
Drittgutachter	Prof. Dr. Leo Wanner

Abstract

The cues that describe emotional conditions are encoded within multiple modalities and fusion of multi-modal information is a natural way to improve the automated recognition of emotions. Throughout many studies, we see traditional fusion approaches in which decisions are synchronously forced for fixed time segments across all considered modalities and generic combination rules are applied. Varying success is reported, sometimes performance is worse than uni-modal classification.

Starting from these premises, this thesis investigates and compares the performance of various synchronous fusion techniques. We enrich the traditional set with custom and emotion adapted fusion algorithms that are tailored towards the affect recognition domain they are used in. These developments enhance recognition quality to a certain degree, but do not solve the sometimes occurring performance problems. To isolate the issue, we conduct a systematic investigation of synchronous fusion techniques on acted and natural data and conclude that the synchronous fusion approach shows a crucial weakness especially on non-acted emotions: The implicit assumption that relevant affective cues happen at the same time across all modalities is only true if emotions are depicted very coherent and clear - which we cannot expect in a natural setting. This implies a switch to asynchronous fusion approaches.

This change can be realized by the application of classification models with memory capabilities (e.g. recurrent neural networks), but these are often data hungry and non-transparent. We consequently present an alternative approach to asynchronous modality treatment: The event-driven fusion strategy, in which modalities decide when to contribute information to the fusion process in the form of affective events. These events can be used to introduce an additional abstraction layer to the recognition process, as provided events do not necessarily need to match the sought target class but can be cues that indicate the final assessment. Furthermore, we will see that the architecture of an event-driven fusion system is well suited for real-time usage and is very tolerant to temporarily missing input from single modalities and is therefore a good choice for affect recognition in the wild. We will demonstrate mentioned capabilities in various comparison and prototype studies and present the application of event-driven fusion strategies in multiple European research projects.

Keywords:
Multi-modal Fusion, Affective Events, Real-time Systems

Zusammenfassung

Hinweise auf emotionale Zustände finden sich in zahlreichen Modalitäten. Multi-modale Fusion ist folglich ein naheliegender Weg, die automatische Erkennung von Emotionen zu verbessern. Studien präsentieren oft Fusionsansätze, in denen Entscheidungen synchron in allen Modalitäten gefordert und diese durch generischer Regeln kombiniert werden. Der berichtete Erfolg schwankt hierbei stark, bis hin zu Fällen, in denen multi-modale Systeme in ihrer Qualität unter uni-modalen Ansätzen rangieren.

Auf Basis dieser Beobachtungen untersucht und vergleicht die vorliegende Arbeit zunächst eine Vielzahl synchroner Fusionsansätze und erweitert die Liste dieser traditionellen Methhoden mit maßgeschneiderten Fusionsalgorithmen zur Emotionserkennung, die darauf abzielen, das vorliegende Domänenwissen bestmöglich auszunutzen. Zwar gelingt es hierdurch, die Qualität der Ergebnisse zu verbessern, das zugrunde liegende Problem löst es jedoch nicht: Eine systematische Untersuchung synchroner Fusionsstrategien auf geschauspielerten und natürlichen Daten legt den Schluss nahe, dass synchrone Ansätze eine grundlegende Schwäche zeigen: Die implizite Annahme, dass relevante Hinweise zum gleichen Zeitpunkt in allen Modalitäten vorliegen, ist nur wahr, wenn die untersuchte Emotion äußerst kohärent und klar dargestellt wird - was jedoch in einem natürlichen Umfeld nicht immer zu erwarten ist. Diese Beobachtung impliziert den Schritt von synchronen zu asynchronen Fusionssystemen.

Dieser Schritt kann durch die Anwendung von Klassifikatoren mit Gedächtnisfunktion realisiert werden (z.B. rekurrente neuronale Netze), diese sind aber meist datenhungrig und untransparent. Folglich beschreiben und evaluieren wir eine Alternative zur asynchronen Verarbeitung von Modalitäten: In einem ereignisbasierten Fusionsansatz entscheiden die Modalitäten, wann Information in Form von emotionalen Ereignissen dem Fusionsprozess zugeführt wird. Dies ermöglicht eine zusätzliche Abstraktionsebene im Erkennungsprozess, da erkannte Ereignisse nicht direkt der gesuchten Zielklasse entsprechen müssen, sondern stattdessen Hinweise auf die finale Einschätzung beinhalten können. Der ereignisbasierten Fusionsansatz unterstützt den Einsatz in Echtzeitsystemen, ist robust gegen den zeitweisen Ausfall einzelner Modalitäten und eignet sich folglich hervorragend für den praktischen Einsatz außerhalb einer kontrollierten Umgebung.

Schlagwörter:
Multi-modale Fusion, Emotionale Ereignisse, Echtzeitsysteme

Danksagung

Zuallererst möchte ich mich bei meiner Betreuerin Prof. Dr. Elisabeth André für die Ermöglichung dieser Dissertation, die stets freundliche sowie konstruktive Unterstützung und die immer angenehme Zusammenarbeit bedanken. Ein besonderer Dank für das Interesse an dieser Arbeit gilt auch meinen beiden Zweitgutachtern Prof. Dr. Björn Schuller und Prof. Dr. Leo Wanner.

Wie wichtig gute Kollegen und eine freundschaftliche Arbeitsatmosphäre sind sollte nie unterschätzt werden. Deshalb geht mein herzlichster Dank für hervorragende Zusammenarbeit, gelungene Projekte und erinnerungswürdige Dienstreisen an Johannes Wagner, Tobias Baur, Dominik Schiller und alle weiteren Mitarbeiter des Lehrstuhls.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In 2006 Polikar [135] presented a comprehensive overview, describing possible approaches to form and evaluate an ensemble of classifiers and the expected benefits of ensemble based systems in decision making. He compared the usage of an ensemble of classification models to several real-life decision making and voting processes and showed the possible advantages over single expert systems. Though this was the author's first contact with multi-classifier systems, it is by far not the first work on ensemble based machine learning. Instead the origins of this field of research are herein dated back to 1979 (Dasarathy and Sheela [37]) when the use of several classifiers to partition an available feature space was first discussed, but it shows a huge number of possible strategies to fuse results of multiple classification systems. However, most considerations in this survey relate to a single source of information, which is processed repeatedly (e.g. by bagging, Breiman [20]) in order to train diverse classifiers. This consequently implies that the described effort of building classification ensembles "only" serves the improvement of classification accuracy. Today, ensemble based fusion techniques play an important role in scenarios where the information that can be used to solve a given classification problem is composed of different data sources that need to be brought together.

In the late 90's the field of *Affective Computing* (Picard [133]) has emerged with the goal to give machines the ability to recognize human emotions. Ever since have research groups recorded and evaluated affective datasets and have tried to decrypt the social signals transported within. This is where a great chance for ensemble based decision making and multi-modal fusion can be found: In human interaction, social signals are generally expressed through multiple channels. Emotions in particular are illustrated by a combination of vocal behaviour, facial expressions, gestures and postures. A generic

approach to the classification of emotions is to choose one type of signal, train a system to extract and recognize preassigned features and cues from it, and finally associate made observations with predefined emotional states. But as humans tend to base and refine their predictions on emotional states on more than one modality, a machine should do so too if possible. Many studies in multi-modal affect recognition have been done by exploiting synergistic combination of different modalities (Wagner *et al.* [172]). Most of the early works focus on fusion of audiovisual information, i.e. combining speech with facial expression. De Silva and Ng [38] as well as Chen *et al.* [30] proposed rule-based fusion methods for a combined analysis of speech and facial expressions. Huang *et al.* [87] used boosting techniques to automatically adapt combination weights for features from audio and video channels. In the work of Busso *et al.* [22], an emotion-specific comparison of modality fusion on different levels has been reported by using an audiovisual database containing four emotions, sadness, anger, happiness, and neutral state, deliberately posed by an actress. Their evaluation proposed that the best fusion method is highly dependant on the application. An advise on a general fusion scheme for affect recognition seems hardly possible.

1.1 Research Questions

The hope behind the idea of including more than one modality for a given emotion classification problem is that additional sources of affective information can contribute to enhanced recognition performance. In addition to the increase in available information, temporal interaction between modalities can be considered and the whole affect recognition system becomes more robust, as modalities that occasionally won't contribute meaningful information can be substituted by other channels. A great amount of fusion systems has consequently been applied within the field of emotion research and computer science. A popular survey performing a meta-analysis on 30 published studies that deal with multi-modal affect recognition was done in 2012 by D'Mello and Kory [48]. It includes a comparison of uni-modal and multi-modal affect detection accuracies and concludes that a majority of systems enhance their performance by applying more than one modality in the recognition process. However, these enhancements are not always given: The effect of multi-modality is in a part of reviewed fusion system either not given or in the worst case a negative effect on overall accuracy emerges. With all the advertised advantages of multi-modal affect detection, how is a negative multi-modal effect possible at all?

Well, first off we need to be careful and have a look on the contributing modalities. We must not expect to fuse approaches that are not working correctly and obtain a better result. In fact it has been proven that the performance of a fusion system is heavily dependent on the recognition accuracy of the individual modalities (Wu *et al.* [182]). This also means that if we e.g. have a good recognition for emotions from facial expressions we can only hope to enhance this uni-modal system with the addition of modalities that contribute with a comparable performance. Given an acceptable quality of all involved modalities, the crucial factor for the accuracy of the fusion system is the manner of combining the individual sources of information. Every modality will introduce errors to the decision process and a good fusion system needs to be built in a way that reduces the influence of these errors instead of potentiating them. Consideration of ways to achieve this goal motivate several questions, studies and suggested solutions that are to be addressed within this thesis:

1. *What defines a good fusion architecture?* As stated above, a substantial factor of the quality of a multi-modal affect recognition system depends on the technology to extract informative features from the single modalities. But the strategy to integrate this information into a coherent decision is of equal importance. The level on which the fusion process should be applied is often discussed and experimented with and possibilities range from early integration on feature level over combination of interim decisions of several modalities and classifiers to late model based and self learning approaches. Most fusion techniques are general rules that can be applied to any combination problem but some affect recognition studies present methodologies that are tailored towards the underlying emotion model. If any of these possibilities is generally advisable needs to be evaluated.

2. *What are the suitable interpretation units for recognizing a given affective state?* The most recent discussions in the field focus on the modelling of temporally shifted occurrences of emotional cues throughout affective channels. While early fusion strategies mainly followed a synchronous strategy to consider all information within a fixed time segment, current systems regard these asynchronous characteristics of emotional manifestations and try to model them within the fusion process. Given a synchronous fusion approach an overarching annotation based mainly on a single modality (such as a spoken sentence or word) is sufficient to define the boundary for analysis of all involved modalities - if however affective cues are expected to happen at shifted points in time we need to find ways to account for these phenomena in annotation, segmentation and fusion modelling.

3. *What is the adequate level of abstraction to describe observed emotional conditions?* The affective states to be recognized are most often labelled either as discrete emotion categories (such as happy, neutral or sad) or on continuous scales (such as pleasure, arousal or dominance). A classifier trained to recognize happiness will most likely look for smiles and grins in the facial modality or signs of laughs and giggles in the voice. Instead of subsuming these cues under one affective category, we could however train classifiers to detect exactly these emotion indicating events and relate their occurrences back to the actually sought emotion. This approach may result in slim cue recognizers with more precise decision boundaries than a classifier dealing with all possible affective cues at once. To this end we will experiment with the concept of affective events. They are defined as short-termed cues indicating an emotional target class and can be asynchronously detected across considered affective channels.

The final goal is to develop a fusion scheme that is able to solve the problems indicated by the given research questions. To achieve this premise, the following requirements have to be met:

- The fusion system should be based on framewise classification to enable the recognition of short-termed affective events and the detection of temporal relations between modalities within an affective episode.
- It needs the capability to model the temporal flow of recognized cues and therefore provide a way to memorize past system states and recognition results.
- In order to regard the (potentially) asynchronous nature of affective channels, we need the fusion scheme to be event-driven. This means the fusion algorithm should act as a client being able to receive and process detected events at any given point in time.
- The developed fusion scheme should be able to handle the additional abstraction layer introduced by the concept of affective events.
- To guarantee practical applicability and the possibility to adapt to a specific affect recognition problem we require modular expandability and a convenient way to handle temporarily missing data.

Figure 1.1 shows a rough schematic of an event-driven fusion algorithm.

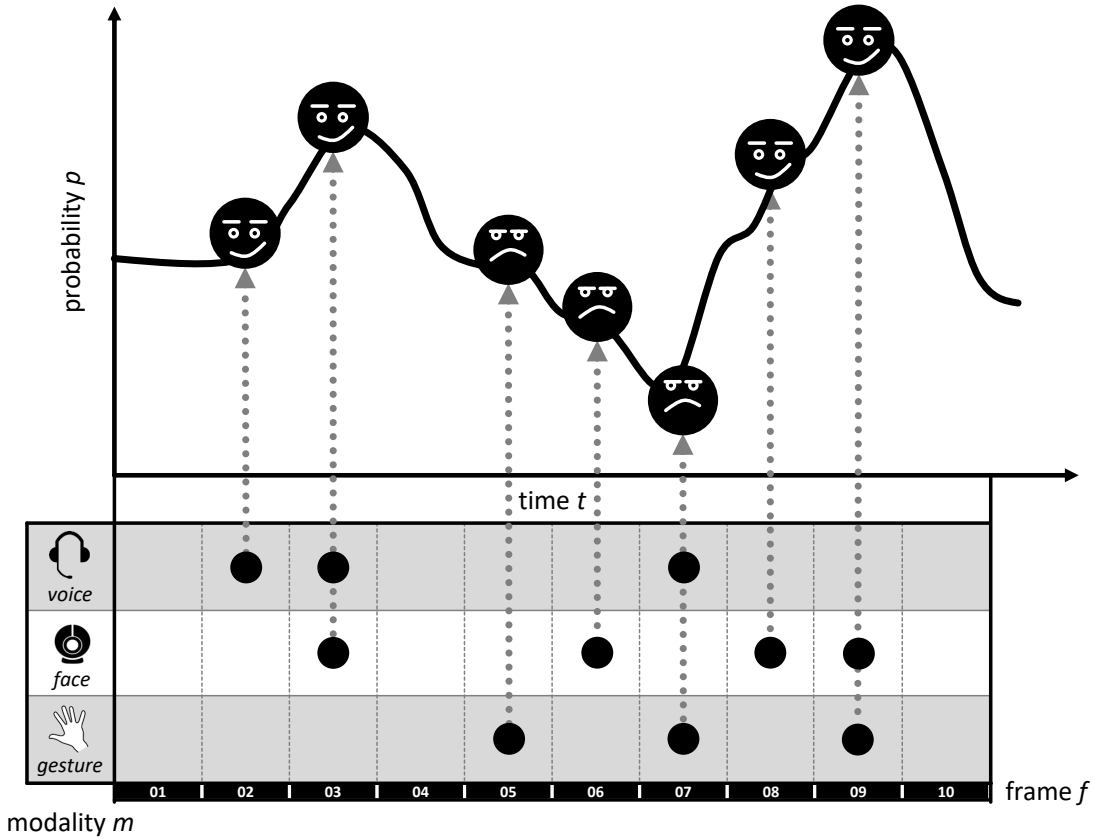


Figure 1.1: Schematic of an event-driven fusion approach.

Events are detected in the single modalities by activity detection and event recognition components (e.g. classifiers) and are processed by the fusion system as affective cues. This way the manifestation of an emotion does not need to be recognisable in all affective channels at once but may be shifted time-wise across modalities. In order to handle this asynchrony, the algorithm keeps track of past events and calculates the momentary probability for a given affective state based on current and preceding events. The initial impact is given by the respective recognition component and can either be a continuous value (e.g. the output of a regression task) or a confidence score given by a classification model. Hereby, the influence of past events decreases over time until they are discarded from the calculation.

1.2 Outline of the Thesis

In the following chapters we will introduce the basic concepts of multi-modal affect recognition, try to identify shortcomings of traditional fusion strategies and consequently work towards concrete implementations of the described event-driven approach:

- In Chapter 2 we introduce the core concepts of affective computing. The affective channels used by humans to (either consciously or unconsciously) communicate emotions are described with respect to their usability in affect recognition systems and the descriptive features that can be derived. Emotion models used to represent and categorize human affect within a recognition system are explained. Afterwards we see how a machine is iteratively taught to understand an emotion via the general affect recognition loop.

- Chapter 3 lays the theoretical foundation for the machine learning and ensemble techniques that are applied within the fusion systems to be discussed. We describe necessary feature extraction techniques as well as used classification models and the methods for their evaluation. Ensemble theory explains the intention, methodology and possible benefits of combining multiple classification models for a given classification problem.

- In Chapter 4 we will see the transition from general ensemble theory to the appliance in multi-modal fusion for affect recognition. Fusion schemes can be implemented at various stages of a multi-modal recognition system and are often categorized by their level of appliance and we will consequently introduce these categories. Afterwards a very comprehensive and recent survey analysing the multi-modal effect on recognition accuracy across a wide range of reported affect recognition systems is reviewed. Here we will conclude that fusion techniques that incorporate the temporal aspects and asynchronous nature of modalities (to which we can count the event-driven fusion schemes to be developed) seem to have an edge over more traditional approaches.

- To start off with practical implementation of fusion architectures, we present first experiments with multi-modal affect recognition in Chapter 5. Though not yet dealing with temporal aspects, we show that custom fusion approaches designed specifically for an emotion recognition problem (emotion-adapted fusion) are able to surpass standard techniques.

- We intensify our investigations on the shortcomings of traditional fusion approaches in Chapter 6 with a systematic comparison of so far introduced fusion schemes. Special attention is given to differences between natural and acted emotions, problems with pre-segmented samples of data, the synchronous consideration of modalities and the impression of interchangeability between standard fusion schemes.

- In Chapter 7 we present our approach to solve the problems identified in preceding experiments: We elaborate the concept of affective events that are asynchronously detected in an ordinal sequence of short frames of multi-modal data. Several event-driven fusion algorithms to process these target-class indicating cues and infer the sought affective states are described.

- Within the next Chapter 8, the developed event-driven fusion systems are evaluated and compared to synchronous standard approaches and state-of-the-art alternatives such as multi-modal neural networks on a corpus featuring naturalistic and non-acted data.

- Practical application of multi-modal fusion in the wild requires the recognition system to be robust against temporarily missing data. Chapter 9 describes the inherent advantages of event-driven fusion schemes in these cases and also provides strategies to implement standard approaches with the capability to handle missing data. We present a mobile event-driven multi-user application evaluated in a real-world setting.

- Chapter 10 gives a short tutorial how to use presented fusion techniques as components in custom affect recognition pipelines within the Social Signal Interpretation framework (SSI) (Wagner [170]). We furthermore demonstrate the successful implementation of multi-modal fusion in several European research projects.

- We conclude the thesis in Chapter 11 with a recap of discovered problems in multi-modal fusion for affect recognition and possible solutions offered by an event-driven approach. We theorise further areas of application in which the additional abstraction layer offered by affective events can be exploited in the future.

Chapter 2

Multi-modal Affect Recognition for Human Computer Interaction

The term human computer interaction (HCI) began to emerge in the late 70's and early 80's and describes a point of intersection between applied psychology, social sciences and information technology. Early studies and theories such as "The Psychology of Human-Computer Interaction" by Card *et al.* [23] suggest that the design of a human computer interface directly influences the human performance in interaction with the machine. Since then HCI has been promoted to be the "visible part of computer science" (Carroll [27]) and is meant to support humans in interaction with computer interfaces of all kinds. HCI motivates system developers to regard the characteristics and behaviour of human users when designing applications (Shneiderman [153, 154]). One crucial part of incorporating the human factor in system interaction is to consider displayed affective states of users, as the depiction of emotions plays a key role in natural human interactions. We experience emotions during each and every moment and are also aware of affective conditions that persons around us are expressing. These affective states are displayed through various channels including facial expressions (Keltner *et al.* [94]), vocal prosody (Juslin and Scherer [91]) or body postures (Dael *et al.* [34]). We learn to interpret these representations during the earliest childhood, as the correct understanding of these multi-modal hints is crucial for a trouble-free communication and interaction with others. Pantic et al. [132] describe a debate whether behavioural signals convey actually felt emotions (Vinciarelli *et al.* [163]) or are more used as a regulator for influencing the process of an interaction (Fridlund [65]). Whether displayed emotions are actually felt or expressed for reasons like conversation techniques, an essential goal of human computer interaction (HCI) is to enable systems to automatically interpret these affective signals,

in order to enhance the interaction between human and machine by augmenting its naturalness (Pantic [132]). In multi-modal affect recognition we consequently aim at developing fusion strategies for reasonably merging the affective information gained from multiple modalities into a common assertion of user emotions. Elaborate ways of fusing multiple modalities are in use throughout many affect recognition studies: (Song *et al.* [158], Zeng *et al.* [186], Wöllmer *et al.* [178],[179], Nicolau *et al.* [126]) combine differing modalities and mostly confirm the assumption that multi–modal fusion generates more accurate affect recognition systems than a uni-modal approach.

2.1 Defining a Ground Truth for Emotions

Before we can start to teach a machine how to interpret cues from multiple modalities and to infer affective states from them, we have to define an affective ground truth on the available data we will need to train recognition systems (Section 2.3). In the case of recordings of emotional states this is a rather difficult task, as labelling observations is not as objective as e.g. tagging a picture with its content. This rather subjective task is sometimes achieved by self-reports of recorded subjects or by a consortium of human experts that try to come to a common conclusion. In order to alleviate and generalize the annotation task, a concept of discrete emotion modelling has to be chosen.

Three terms are commonly used in emotion research: Feelings, emotions and affect (Damasio [36]). Feelings often describe an internal mental state that is not depicted in bodily expressions whereas the term emotions is commonly used to describe the physiological manifestation of these feelings. Affect subsumes both phenomena under a general term, describing the cause as well as the portrayal of human emotions. Within the field of automatic affect recognition more interest is given to the recognition of affective states as to the psychological reasons for their occurrences, so the terms affect and emotion are mainly used as synonyms. These considerations already hint to the fact that affective states are very abstract concepts, describing a vast amount of psychological states. These conditions are too numerous to use them directly for recognition tasks, so they have to be integrated into quantifiable categories of emotions. A discrete emotion model is necessary to define target emotions, so that the recognition system is able to understand the problem to be solved. Moreover, the emotion model supports a convergent labelling process. Such procedures narrow the amount of identifiable feelings and group the wide field of possible individual emotions into a small amount of discrete emotion-classes. Categorical and dimensional models are the two most prevalent approaches to conceptualize human emotions.

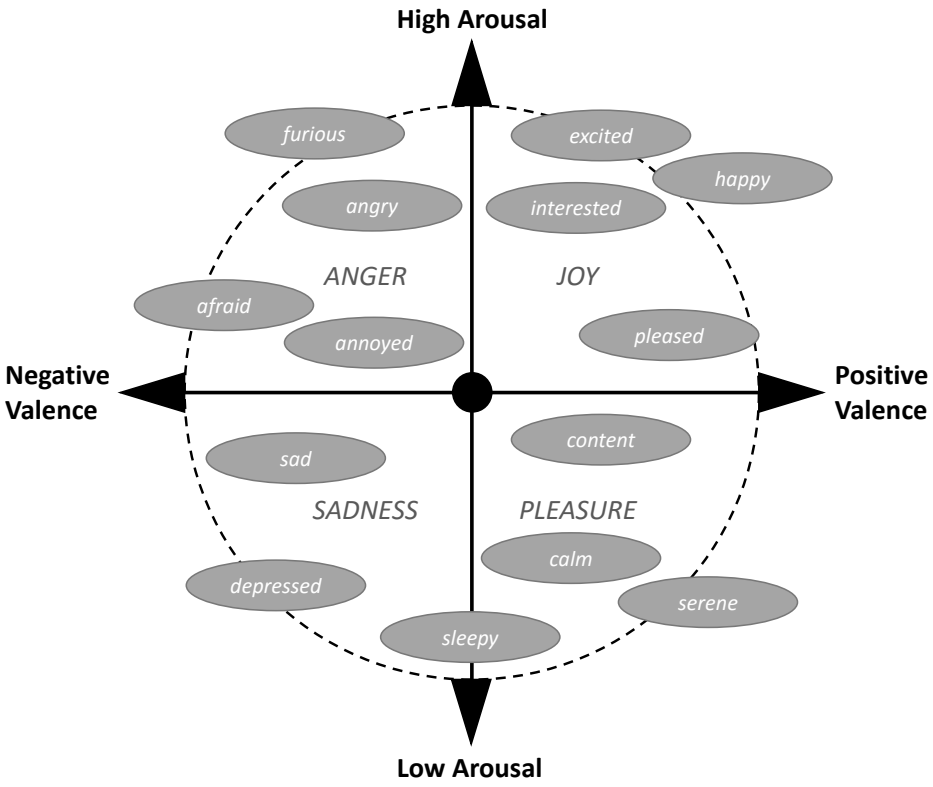


Figure 2.1: Dimensional arousal and valence based emotion model and derived discrete emotion categories.

A categorical model subsumes emotions under discrete categories like happiness, sadness, surprise or anger. They are mostly labelled by external specialists. In early research, Ekman has tried to define sets of basic emotions, which are universally valid among all humans (Ekman [57]). There is a common understanding of these discrete categories, as terms describing the emotion classes are taken from common language. However, this approach could be restricting, as many blended feelings and emotions cannot adequately be described by the chosen categories. Selection of some particular expressions can not be expected to cover a broad range of emotional states and could suffer from randomness.

Another way of describing emotions is to attach the experienced stimuli to continuous scales within dimensional models. Mehrabian [118] suggests to characterize emotions along three axes, which he defines as pleasure, arousal and dominance. Lang *et al.* [107] propose arousal and valence as measurements. These representations are less intuitive but allow continuous blending between affective states. They describe multiple aspects of an emotion, the combination of stimuli's alignments on these scales defines single emotions (Figure 2.1). In the case of the arousal-valence model, which is the most commonly used

dimensional model for affect recognition, the valence scale describes the pleasantness of a given emotion. A positive valence value indicates an enjoyable emotion such as joy or pleasure. Negative values are associated with unpleasant emotions like sadness and fear. This designation is complemented by the arousal scale which measures the agitation level of an emotion.

Categorical as well as dimensional models are simplified and synthetic describtions of human affect and are not able to cover all of the included aspects. They are however useful and needed to model emotions as concepts to be presented to a machine. In addition to the categorical and dimensional models, appraisal models are in use to simulate the emotional behaviors of virtual agents in dialogue systems (Bee *et al.* [10]). Among the most known examples are EMA (Gratch and Marsella [75]) and ALMA (Gebhard [69]). There are only a few approaches that rely on appraisal models for affect recognition tasks (see Bosma and André [18] for an early example). More recent work by Mortillaro *et al.* [122] suggests to infer emotions based on autonomic symptoms and motor behaviour as appraisal results (see also Section 7.2.1). A first approach into this direction has been made by Soleymani [156] with the detection of appraisal components, such as novelty, from facial expressions.

2.2 Affective Channels

Given appropriate sensor technologies, signal processing methods can be applied to extract descriptive features from sensor data over fixed time frames. Classification models are able to use these features for interpretation of recorded signals by learning a mapping between observed features and discrete labels, e.g. affective states. A classification model tries to associate an unknown sample with a pre-defined category the mapping fits best. In uni-modal affect classification, features of one social channel, such as the observed vocal properties, are used to make assumptions about the current emotional condition of a user (Eyben *et al.* [60]). But as the cues that describe emotional conditions are indeed encoded within multiple modalities, the classification process should incorporate as much multi-modal information as possible from multiple channels (Zeng *et al.* [187]).

2.2.1 Paralinguistics

Speech is the main channel through which humans exchange information. Human language encodes emotional information in two different ways - what is said and how

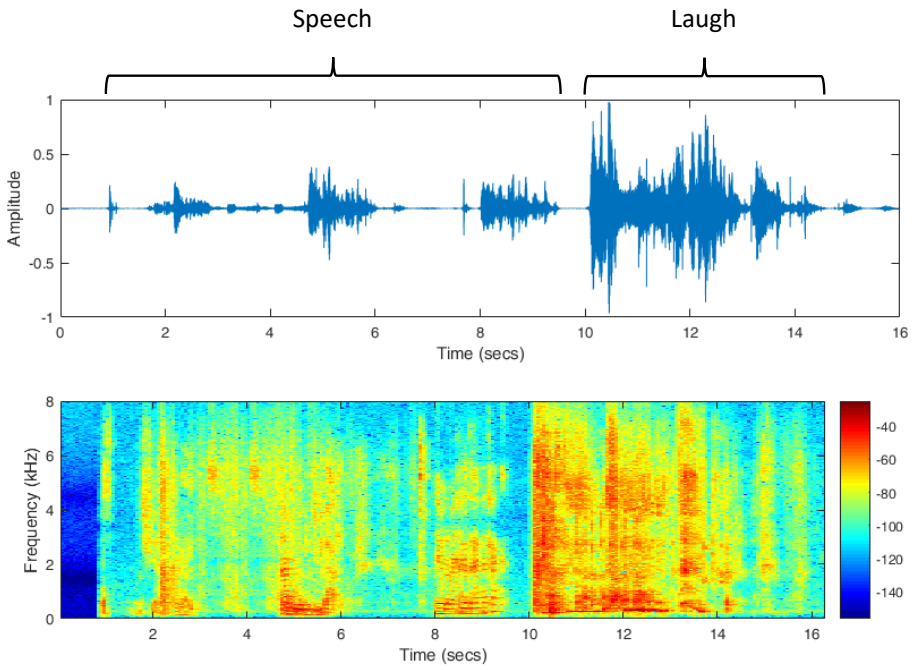


Figure 2.2: Raw audio signal and spectrogram of a person talking normally and then bursting into an affective laugh. Though in this case differences between the characteristics of the two signal parts can be perceived by the eye, the list of proposed paralinguistic features actually suited for automated extraction and affect recognition is long.

it is said (Wagner *et al.* [172]). The transmitted content can be split into two parts, a semantic and a paralinguistic message. The importance of the paralinguistic message is evaluated by Mehrabian [119, 121]: He proposes that a feeling of liking or disliking depends to 55% on facial expressions and to 38% on paralinguistic cues, but only to 7% on the spoken words. Though the so called *7%-38%-55%-rule* is often misinterpreted as general rule of thumb (Mehrabian clarified that this only holds for the confined boundaries of his attitude study), we can conclude an important role of prosodic information for affect recognition.

Paralinguistic aspects include acoustic features like pitch, loudness or speech rate and based on these features, it is possible to assign emotional labels to utterances. Figure 2.2 shows the audio signal of a person's normal speech leading into an affective burst of laugh. A classical paralinguistic affect recognition task would now be to automatically differentiate these parts of the signal. Though differences between the characteristics of the two signal parts can in this case be perceived by the eye, the list of proposed paralinguistic features actually suited for automated extraction and emotion recognition is long. A co-operation of different sites under the name CEICES (Combining Efforts for

Improving Automatic Classification of Emotional user States) has carried out experiments based on a pool of more than 4000 features including acoustic and linguistic feature types (Batliner *et al.* [7]). Results for the individual groups, as well as a combined set, have led to the following assumptions: Among acoustic features duration and energy seem to be most relevant, while voice quality showed less impact. Yet, no single group outperformed the pool of all acoustic features. Several tools for automatic feature calculation are in active use and offer a convenient way to install well sorted and tested feature sets in practical applications. The most prominent representatives include the Munich Versatile and Fast Open-Source Audio Feature Extractor (openSMILE) (Eyben *et al.* [59]) (which includes the well known Geneva minimalistic acoustic parameter set (GeMAPS) (Eyben *et al.* [62])), Praat (Boersma and Weenink [15],[16]) and EmoVoice (Vogt *et al.* [166]).

Overall the automated recognition of affective states turns out to be a challenging task, as the variations in features accross various styles of speaking are large whilst machine learning algorithms in general need to be robust against outliers (Trigeoris *et al.* [160]). Long Short-Term Memory (LSTM) networks (Section 3.2.4) are currently used in the most promising studies for paralinguistic affect recognition (Brückner and Schuller [21], Wöllmer *et al.* [179]).

2.2.2 Facial Expressions

Facial expressions can be called the most expressive transmitter of human emotions. Concentration and training would be needed to mask the depiction of affect in one's face reliably. But even then it takes a certain amount of time to control the muscle reactions that were triggered by the underlying affective state and the correct emotion is expressed during this short period (Ekman [54]).

Automatic facial behaviour analysis has classically focused on the recognition of universal facial expressions and Action Units (AU) - atomic movements in the face caused by the activation of one or more facial muscles. The research is mainly motivated by well-known studies of the psychologists Ekman and Friesen [58]. During early studies with an isolated population from New Guinea Ekman showed that there exist 6 universal and culture independent emotions (anger, happiness, fear, surprise, sadness, and disgust) and that each of them has a corresponding prototypical facial expression. By now it has been demonstrated (Du *et al.* [50]) that people can perform many other non-basic expressions representing e.g. contempt, embarrassment or concentration and that combinations of these expressions are commonly found in every-day life scenarios. In 1978 Ekman and Friesen introduced the Facial Action Coding System (FACS) (Ekman and Friesen

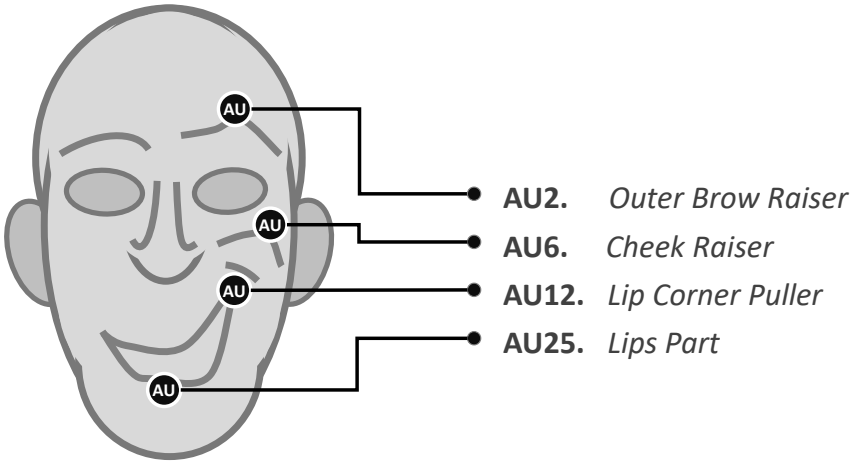


Figure 2.3: Facial action units and their activation in prototypical emotions.

et al. [55]). The coding system defines a set of Action Units for possible movements of the human face's musculature. Any possible facial expression can be inferred from a concrete combination of Action Units. This way it is possible for trained coding experts to annotate emotions by referencing the state of these units. These configurations can directly be used as features for machine learning approaches, which makes automatic Action Unit recognition one of the most interesting problems in facial behaviour analysis. Of special interest for affect recognition is the further development of FACS into FACSAID (Facial Action Coding System Affect Interpretation Dictionary) (Ekman *et al.* [56]) which consider only emotion-related facial actions.

There is a wide array of software (partly open-source) for the automated analysis of facial expressions which provide easy access to face tracking, facial landmark estimation, action unit assessment and sometimes even emotional labelling. Solutions often used in research studies include the SHORE library (Ruf *et al.* [142]), Openface (Baltrušaitis *et al.* [5]) and eMax (Almaev and Valstar [1]).

2.2.3 Gestures and Postures

The actions and positions of of body, head and limbs - also referred to as *kinesics* - have for a rather long time not been the main focus of coordinated emotion research (Harrigan [79]), resulting in a lack of clearly defined coding systems such as the Facial Action Coding System (FACS) (Section 2.2.2) for facial expressions. Although early work like Caridakis *et al.* [24] describes the possibilities of expressivity features (mainly calculated on arm and head movements extracted from video sequences), the main assumption was that body movement only shows intensity of emotions. Therefore

emotion communication research mainly focused on the face and voice as assumably more expressive modalities (Dael *et al.* [33]). However, in recent years studies have shown that dynamic body movement and gestures as well as static postures (Atkinson *et al.* [3, 4]) actually convey affective states of a monitored person. Affective concepts such as a person's current expressivity (Caridakis *et al.* [26]) or engagement in a conversation (Baur *et al.* [9]) can reliably described by suitable movement features (Hartmann *et al.* [80]) and body orientations and postures. The trend to more accurately analyse gestures and postures as a mean to affect recognition was most likely positively affected by technical advancements like the Microsoft Kinect™, that make whole body tracking a more accessible approach. By 2012 a comprehensive survey by D'Mello and Kory [48] investigated 30 current multi-modal affect recognition systems and stated that almost a third of these systems were by then using information from some form of body movement, postures and gestures as a source for emotion assessments.

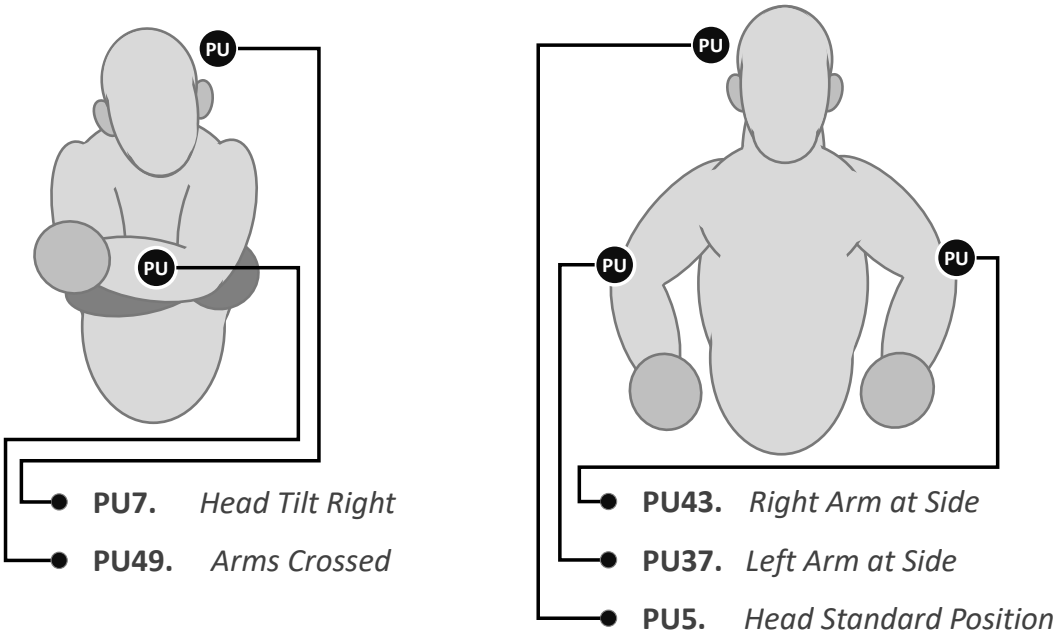


Figure 2.4: The Body Action and Posture System. Posture units (PU) describe the general alignments of trunk, head and limbs to a resting configuration.

A logical consequence of this rising interest in kinesics is the effort to develop a reliable coding system for body movement in emotion expression, just like the by now available standards for facial expressions. For this reason Dael *et al.* [34] describe the Body Action and Posture System: The system distinguishes body posture units and body action units (Harrigan [79]) (Figure 2.4) with the first representing the general alignments of trunk, head and limbs to a resting configuration (e.g. arms crossed) and the latter one describing a local and short termed movement of head or arms (e.g. pointing gesture).

2.2.4 Physiological Responses

In contrast to the so far mentioned modalities, physiological responses are not actively used by humans to express and depict emotional states. Instead they are more unconscious reactions to affect that are hard to control by untrained subjects. A person's physiological reactions are hard to mask or control, as they are regulated by the autonomous nervous system. Physiological signals are a valuable source of information for affect recognition tasks, they are always present and can therefore be captured continuously.

- **Heart Rate and Blood Volume Pulse**
  Electrocardiography (ECG) describes the recording of electrical heart activity with electrodes placed on the skin. An ECG signal consists of repeated QRS complexes that are continuously created by the electrical impulses that occur during the depolarization of the right and left ventricles of the beating human heart. The easiest way to automatically detect a heartbeat within an ECG signal is to locate the R-spike contained as high amplitude within each and every QRS complex. By counting the R-spikes over time, the heart rate (HR) can be calculated as heartbeats per minute. This measurement together with heart rate variability (HRV), that describes variations in the time intervals between the single R-spikes is commonly accepted indicator in social communication (Quintana *et al.* [137]) and affect recognition - mainly applied to arousal related emotional states, e.g. relaxed conditions versus joy and fear (Valderas *et al.* [161]). The measurement of blood volume pulse (BVP) is an alternative to electrocardiography. The used sensor is a photoplethysmograph which measures infra-red light reflections from the skin. When blood is pumped from a heartbeat, the engorgement widens the veins under the skin and they reflect more light. This way the heart rate can be estimated. The sensor can be worn on the finger and does not need moistening or gel and is therefore more easy to use than ECG. On the other hand the signal is not as precise as an ECG signal and often subject to artefacts, making the calculation of HRV features difficult.

- **Skin Conductivity**
  The skin is constantly undergoing a variation of electrical conductivity. Early studies on emotion have already confirmed that the magnitude of skin conductivity (SC) and affect (mainly on the arousal scope) are closely associated (McCurdy [115]). A skin conductivity sensor is considered to measure the activity of sweat glands and the skin's pore size. It records the skin's ability to conduct electricity via a small voltage applied to the skin and the skin's current resistance. The

SC signal is characterized by a tonic (i.e. skin conductance level) and a phasic component (i.e. skin conductance response). The skin conductance level is a slowly varying baseline, while the skin conductance response is a relatively fast rise in the signal related to an external stimulus. This rise typically emerges within a window of up to four seconds after the stimulus is given. So in order to obtain features from the signal that can reliably be related to affective experiences for emotion recognition studies (Picard *et al.* [134], Wagner *et al.* [167]), the tonic baseline is usually removed and statistical measures or peak detection algorithms are applied to the phasic component.

- **Muscle Activity**
  Electromyography (EMG) describes a diagnostic procedure which is able to measure electrical activity caused by muscle contraction with the help of an electromyograph. The stronger a contraction, the more muscle fibres are used and this directly manifests in an amplitude of proportional intensity within the signal measured by the corresponding electrode. Of special interest for psychophysiology is the facial electromyography in which electrodes are placed on facial muscles that indicate valence related emotional expressions e.g. smiles or frowns. Fridlund *et al.* [66] presented a very early pattern classification system to recognize happy, sad, anger and fear poses from facial EMG. More present studies often focus on facial reactions in human-human (Dimberg *et al.* [46]) or human-agent (Weyers *et al.* [176]) interactions and mimicry effects (Hess and Blairy [82]). EMG electrodes placed on frontal muscles or the neck are clinical indicators of arousal and statistical features gained from these observations can be used for automatic emotion classification (Kim and André [96]).

- **Respiration**
  Breathing activity or respiration (RSP) is typically measured by voltage changes caused by stretching within a respiration belt. This way the rate of respiration as well as the depth of breaths can be captured. An even more precise respiration measurement is the observation of gas exchange, but sensor devices are mostly too intrusive to be concerned for HCI purposes. Boiten *et al.* [17] give an overview which respiratory patterns suggest certain emotional states. An increase in breath intervals and deep breathing are reported to indicate states of high arousal such as stress or fear, while relaxed states often show opposing patterns. The length of inhalation and exhalation within a breath cycle can also be assessed for a very detailed analysis. In general, RSP is not the most accurate modality for affect recognition by itself, but is often used as additional source of physiological

information or to classify physical activities like talking or laughing.

- **Body Temperature**
  Body temperature (TEMP) is a more exotic physiological measurement outside of clinical purposes which is not often used in emotion recognition approaches. The signal can be captured by very sensitive temperature sensors from the skin or via infra-red thermal cameras. Given a healthy user-state, the subtle changes in body temperature are related to the flow of blood within the body and is therefore qualified to be considered for affect recognition in a similar sense to ECG or BVP. Puri *et al.* [136] relate body (i.e. forehead) temperature to stress and frustration and try to recognize it in real-time with a thermal camera. Though of course the information gained from the signal is much more general than heart rate features, the advantage of such an approach in comparison to the use of an ECG or BVP device is the non-intrusive nature of the sensor, which in contrast to most physiological sensors does not necessarily need to be in contact with the skin.

- **Pupil Dilation**
  An eye's pupil diameter is able to narrow or widen as physiological response. The pupil dilation (PD) is the widening response and is expected to increase under (mentally) stressful conditions. To use the signal for affect recognition, the unavoidable artefacts introduced by blinking should be removed via interpolation. Afterwards the signal amplitude can be directly used as stress indicator (Barreto *et al.* [6]). In [129] and [171] Omedas *et al.* describe a sensing architecture to recognise conscious and unconscious user reactions within an immersive mixed reality space. Among other physiological sensors within the framework, a small head mounted camera is used to capture the PD signal and map it to the cognitive workload of the person within the mixed reality environment.

- **Electroencephalography**
  Electroencephalography (EEG) describes the normally non-invasive measuring of electrical brain activity. With the help of electrodes placed around the head, voltage fluctuations are measured over time and typically examined for spectral content. In a clinical setting, high quality but expensive devices are used in combination with intrusive skull caps for exact electrode placement. This approach delivers clean and precise EEG data and is mostly used for medical diagnosis of disorders such as epilepsy, sleep disorders and strokes. EEG signals can also be used for affect recognition purposes. Pilot studies and emotional database research such as [99] and [100] (Koelstra) report the possibility to correlate states of high and low

valence, arousal and liking scores with observed EEG frequency bands. Due to the stationary setting and the low resistance to disturbances of the recorded signals, affective states are mostly elicited by emotional video and audio clips. Recent years have seen a number of low-budget, mobile EEG devices that promise a more flexible handling, opening more fields of application. However, one needs to be aware that the recorded signals are very coarse and to a high amount depict facial muscle activities instead of actual neural oscillations.

Physiological responses are less often found in multi-modal affect recognition studies than the so far mentioned modalities. This may be due to the more challenging recording setups needed to capture the sometimes obtrusive physiological sensors. However, the main reason is probably the fact that physiological affective reactions can (for the most part) not be acted. That means that for recording these reactions, the emotions have to be reliably elicited during the data collection. Compared to the recording of an acted audio-visual corpus, this is a complex task requiring psychological finesse and therefore the number of corpora featuring physiological affective data is substantially lower. Nevertheless, a lot of research is done in the field: Early on Picard *et al.* [134] showed that affective states can automatically be recognized by heart rate, skin conductivity, muscle activity, respiration and body temperature. Due to the aforementioned difficulties, Picard recorded eight different emotional states of a single subject over the course of multiple weeks, using the Clynes protocol (Clynes and Menuhin [32]) and the corresponding computer controlled prompting system. The subject-dependent off-line study was able to reach classification accuracy of about 81% for an eight-class classification problem, encouraging future affect recognition efforts to include physiological signals. In the following years Nasoz *et al.* [124] used video clips for affect elicitation, while Wagner *et al.* [167] relied on stimulating music clips to record affective corpora featuring a comprehensive set of physiological responses. Since these early studies on the linking of physiological responses and emotion recognition more and more studies have incorporated these measurements in affect recognition systems. Toolboxes such as the Augsburg Biosignal Toolbox[1] (AuBT) (Wagner *et al.* [167]) and the Toolbox for Emotional feAture extraction from Physiological signals (TEAP) (Soleymani *et al.* [157]) have been established to ease the task of handling the various physiological channels and measurements.

[1] `http://mm-werkstatt.informatik.uni-augsburg.de/files/project_content/33/219_AuBTGuide.pdf`

2.3 The Affect Recognition Loop

Having chosen a way to represent the emotional concepts, a machine can learn to categorize observed emotional cues. A recognition system needs to be trained with a set of representative samples. Collection of these samples is usually done during separate recording sessions in which users are either asked to show certain emotional states or interact with a system that is designed to induce the desired behavioural reactions. The collected data needs to be observed by annotators with the aim to give labels the recorded user actions. The following automatic interpretation of affective signals in affect recognition systems can generally be divided into to five successive tasks (Figure 2.5).

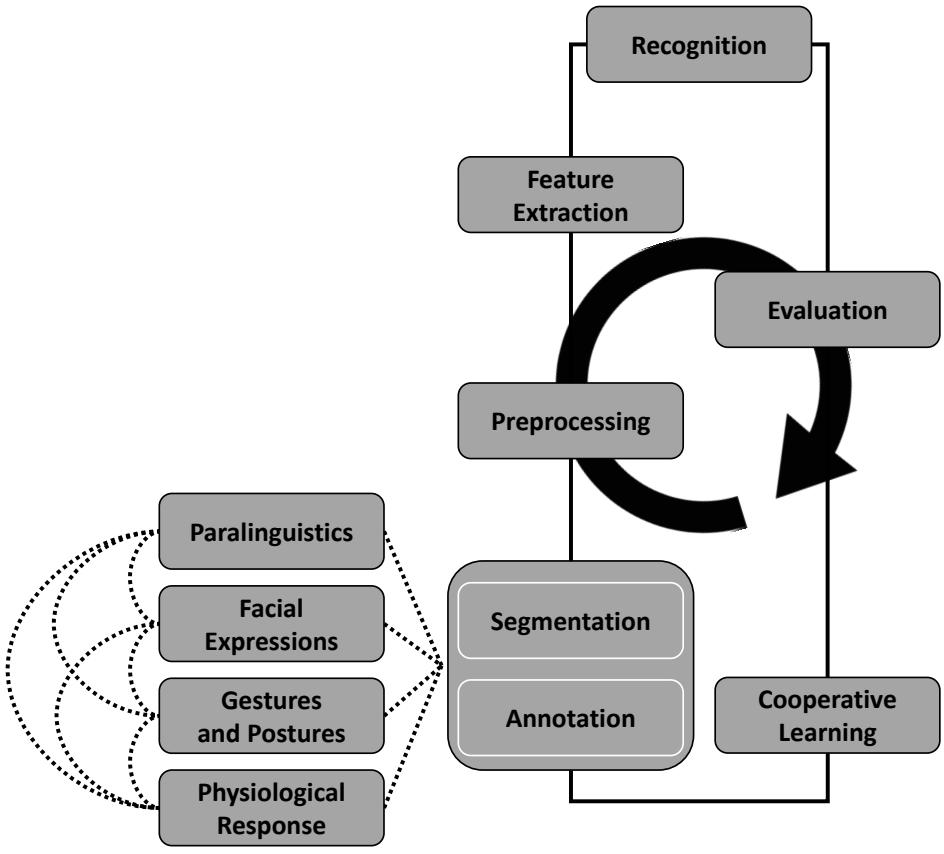


Figure 2.5: Affect recognition loop containing annotation, preprocessing, feature extraction and classification steps.

- Data annotation and segmentation is the process of determining on- and offsets of relevant episodes in observed channels. The units to be identified can be syntactically identified, triggered by activity detection algorithms or defined by

slicing signals into frames of fixed size. Resulting data segments are the so called samples. These are identified with an emotional categorization resulting from the chosen emotion model (Section 2.1).

- Preprocessing steps are necessary filter and/or transformation actions that can enhance the quality of captured data or translate signals into a more informative depiction.

- Feature extraction reduces a sample's raw (or preprocessed) information to a set of coefficients. These features are meant to include the relevant information in descriptive numbers that help classifying the content.

- Recognition describes the process of mapping of samples (and/or respective feature vectors) onto a set of discrete categories (classification) or continuous values (regression).

- Evaluation against the affective ground truth is needed in order to estimate the quality of the developed recognition model. If the results are unsatisfactory, we can re-access, modify and re-iterate the loop at any given pre-step.

- Cooperative learning is a novel approach to augment and extend the annotation of available data. Once a satisfactory recognition performance is reached, the machine's ability of rapid computation can be used to automatically label additional data in cooperation with manual annotation efforts of human raters (Zhang *et al.* [188]).

Chapter 3

The Basics of Ensemble-based Machine Learning

We have described the theoretical affect recognition procedure in Chapter 2.3. But in order to realize the single steps needed in an actual affect recognition system one has to take a closer look at the necessary machine learning basics. Therefore we will in this chapter introduce the basics of feature extraction (Section 3.1) on a short- and long-term scope, a number of often used statistical classification models and ways to evaluate (Section 3.2) their quality. As we are aiming for multi-modal emotion analysis, assumptions concerning ensemble theory and diversity (Section 3.3) need to be introduced as well. The list of possible approaches to these signal processing and machine learning techniques is vast, discussed algorithms are chosen to cover the algorithms used in the studies to be presented in the following chapters.

3.1 Descriptive Features

Features are representative values that are meant to describe the properties of the underlying signal in a way that is well fitted to differentiate signal samples based on the given classification problem. Therefore features are often computed after a preprocessing phase during which additional properties of the signals are carved out and unwanted aspects are suppressed. We have seen examples of suitable features for various modality and signal types in Chapter 2.2. A typical preprocessing measure would be the removal of the tonic baseline in the skin conductivity signal (Chapter 2.2.4). This step is necessary because informative statistical features and peak detection algorithms can only be reliably applied to the phasic component of the signal.

The extraction of descriptive features brings the raw signals into a compact and descriptive form that is required by most classification models. The recently rising interest in neural networks and a steady increase in computational power however encourage studies that try to skip over the feature extraction step. Most recently Trigeorgis *et al.* [160] experimented with a combination of convolutional and long short-term memory neural networks in order to automatically learn the best representation of paralinguistic properties from raw signal values. But although the proposed end-to-end speech emotion recognition system achieved good results in comparison to a traditional feature extraction approach, the majority of classification systems will in the foreseeable future still depend on well designed and descriptive features.

3.1.1 Long and Short Term Observations

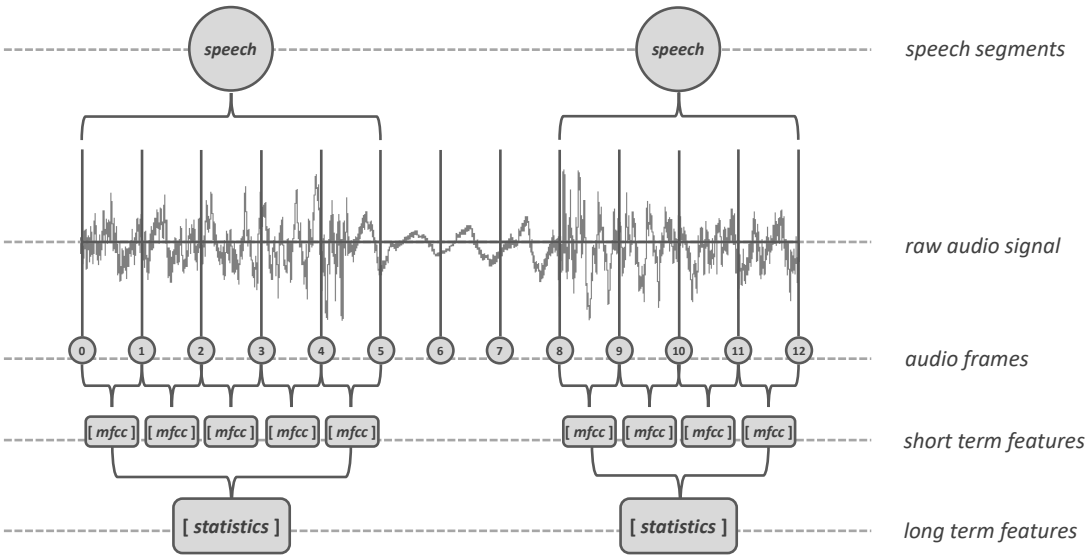


Figure 3.1: Simplified scheme of a feature extraction procedure on raw audio data. Mel frequency cepstral coefficients (MFFC) are extracted over relatively short frames of the signal (short-term features). Statistics such as mean, median and variance are then calculated over the MFFC values of several frames (long-term features).

Depending on whether the features are extracted on a small running window of fixed size or for longer chunks of variable length, we denote them as short- or long-term features. Figure 3.1 shows a scheme of short- and long-term feature calculation. In order to better describe the information contained in the raw audio signal, Mel frequency cepstral coefficients (MFFC) (often used in feature sets for emotion recognition from

voice (Lalitha *et al.* [106])) are calculated over short frames of the audio signal. To produce an assertion over longer periods of time (up to a whole sample of audio, e.g. a spoken word or sentence given by voice activity detection or manual segmentation) statistics over the MFCC values are calculated for the respective segment. The simplified scheme shows non-overlapping frames while in real applications typically an overlapping sliding window of fixed length is shifted over the signal.

3.1.2 Feature Reduction

The goal of feature reduction is to reduce the dimension of the feature space by eliminating redundant information that disturbs the training of classification models. This can either be achieved manually by expert knowledge or in an automated way. In the latter (and very popular) case, one can rely on algorithms for dimensionality reduction, which actively transforms the features into a more streamlined representation, or automatic feature selection, in which original feature values remain untouched.

Often used for dimensionality reduction is the principal component analysis (PCA) (Wold *et al.* [177]). Generally said, PCA simplifies a data matrix by converting presumably correlated variables (features) into a set of uncorrelated variables, the so called principal components. The technique is often applied to transform high dimensional feature spaces into low dimensional variable sets that can be used for visualisation and outlier detection. The transformation parameters are highly dependant on the characteristics of known data and dimensionality reduction is therefore mostly used for data exploration purposes. The synthetic features received by dimensionality reduction do furthermore not consider the suitability for a given classification task, consequently feature selection is the more convenient approach that is mostly chosen for recognition systems.

For automated feature selection a whole set of calculated features is fed into a selection algorithm, that tries to find a minimal subset of the most representative and significant features. Two major advantages come along with this reduction of feature dimensionality: A reduced feature set hugely decreases computational efforts for the upcoming classification steps and a remarkable increase in overall classification accuracy can usually be observed. The drawback is a possible decrease in generalisability of the resulting classification model. The chosen set of selected features may work well on the given data, unseen samples may show differing characteristics and could possibly benefit from other available but not selected features.

The most promising of these algorithms, regarding improvements in classifier performance, are greedy search strategies, that tend to require large amounts of calculation

time. A large number of algorithms have been developed and discussed over time (Guyon and Elisseeff [76]), but two specific approaches (with several variations) have established great popularity: sequential forward selection (SFS) and sequential backward search (SBS). While performing forward selection, variables are included into growing subsets step by step, whereas backward search starts with the complete set of variables and progressively discards the variables that are rated to be the least promising ones. Both algorithms are expected to find good, but not optimal subsets of features, whilst SFS should be more effective on the computational side but potentially generates slightly inferior selections, as the importance of variables is not evaluated in the context of other variables that are not yet included in the set. An often applied alternative to the described brute-force selection algorithms is the correlation-based feature subset selection (CFS). This selection technique searches for subsets of features that show an optimal relation of being highly correlated with the class whilst correlating with each other as little as possible [1].

3.2 Classification Models

In statistical analysis, a classification model is an algorithm that tries to map unknown data instances onto a set of discrete categories or continuous values. Classification is an example of pattern recognition. Figure 3.2 shows a simplified schematic representation of a classification system. During the supervised learning process, a decision boundary is derived from a set of training samples which are given to the classification model to determine the internal variables on which a classification will depend. A training sample consists of a set of numerical representatives (features) that describe the sample. These features are accompanied by a a discrete label (or a continuous value in case of regression) that holds the ground truth for the respective sample. When confronted with unseen data, the trained classification model will apply the developed decision boundary to categorize the new input into known classes. Most experiments presented in this thesis apply discrete classification models.

Regression on the other hand tries to predict continuous-valued output from given descriptive features. Let us take the possible ground truth definitions (Chapter 2.1) for emotions as example: While a discrete classification model would be well suited to categorize emotions into discrete labels, we face a regression problem when we try to assess the correct position of an observed affective state on the axis of a dimensional

[1] Algorithm available within the Weka toolbox `http://www.cs.waikato.ac.nz/ml/weka/`

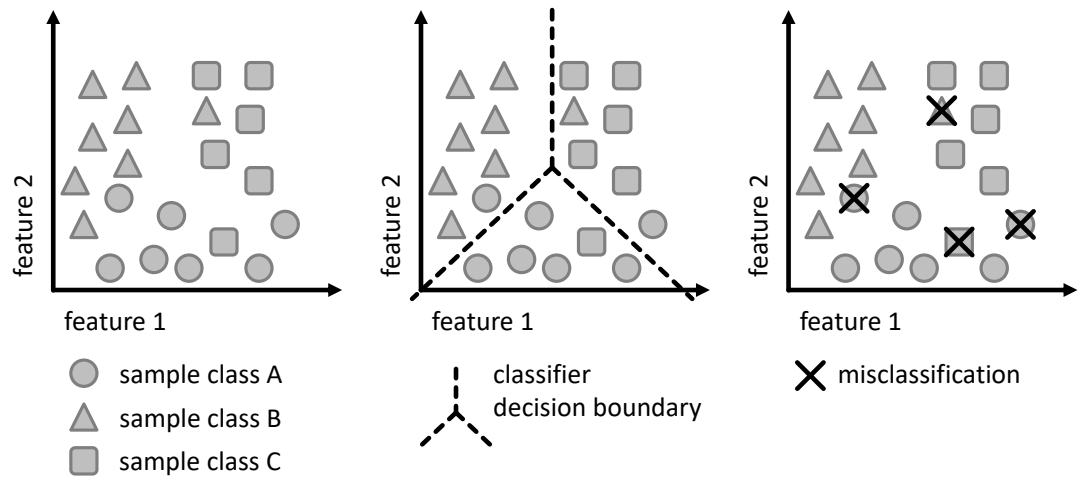


Figure 3.2: Schematic representation of a classification process. A classifier derives its decision boundaries from training data. the decision process is unlikely to be perfect, leading to a misclassification rate which defines the quality of the classification model.

model (e.g. the valence and arousal scales). In this case we expect a continuous value as classification result. Figure 3.3 shows the principle of a regression task. Based on features of the training samples, we estimate a regression function with minimized regression errors that can be used to calculate a continuous regression output for unknown samples.

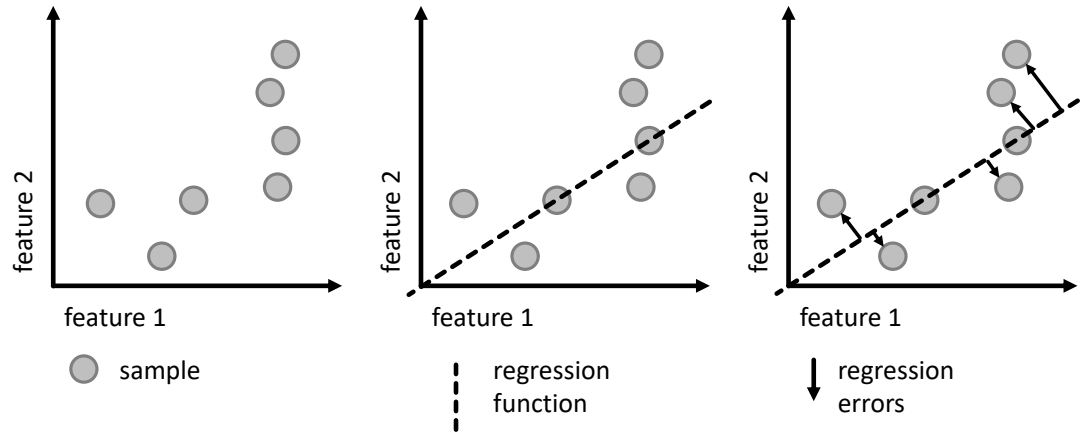


Figure 3.3: Schematic representation of a regression process. Based on features of training samples, we estimate a regression function. Hereby, regression errors need to be minimized. The function is used to predict the continuous regression output for unknown samples.

3.2.1 Naive Bayes

The naive Bayes (NB) classification model is based on the Bayes Theorem and makes the naive assumption of observed feature variables being mutually independent. It is a very fast and memory-efficient model and often performs well in practical applications. The classification algorithm can be defined as follows (Han *et al.* [78]):

Let $X = (x_1, ..., x_n)$ denote an unknown sample with n-dimensional feature vector and $C_1, ..., C_k$ the k classes sample X can be associated with. Sample X is classified as member of class C_i if the following requirement for $1 \leq j \leq k, j \neq i$ is met:

$$P(C_i \mid X) > P(C_j \mid X) \tag{3.1}$$

Following the Bayes Theorem the probability of sample X belonging to class C_i can be calculated as:

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \tag{3.2}$$

Class C_i for which $P(C_i \mid X)$ is maximized is called the *maximum-a-posteriori* estimation.

Given dataset D with $|C_{i,D}|$ being the number of instances in the dataset belonging to class C_i and $|D|$ the absolute number of instances, the *a-priori* probability for each class is defined as:

$$P(C_i) = \frac{|C_{i,D}|}{|D|} \tag{3.3}$$

Under the naive assumption of observed feature variables being independent, $P(X \mid C_i)$ is calculated as:

$$P(X \mid C_i) = \prod_{l=1}^{m} P(x_l \mid C_i) \tag{3.4}$$

In case of continuous feature variables the conditional probability $P(x_l \mid C_i)$ is estimated by a Gaussian distribution with mean score μ and standard deviation σ:

$$P(x_l \mid C_i) = \frac{1}{\sqrt{2\pi\sigma_{C_i}}} exp^{-\frac{(x_l - \mu_{C_i})^2}{2\sigma_{C_i}^2}} \tag{3.5}$$

In order to classify sample X we calculate $P(X \mid C_i)P(C_i)$ for each class C_i. The sample is categorized as class C_i if

$$P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j) \tag{3.6}$$

is valid for all j with $1 \leq j \leq k, j \neq i$, which means it is the maximum.

It is possible to adjust the Bayes formula to handle regression problems by constructing a proximity function for $P(V \mid X)$, where V is the continuous target value.

The rather simple NB classifier has been used with comparable results to more sophisticated approaches such as support vector machines (Section 3.2.2) in affect recognition studies (Vogt and André [164, 165]). Given the low demands on training data size and fast computation times, it is well suited for exploratory tasks and we have consequently applied the model numerous times within our first fusion experiments presented in this thesis (Chapter 5).

3.2.2 Support Vector Machines

The support vector machine (SVM) approach to discrete classification is to separate instances of different classes by means of a hyperplane with maximum margin (Chang and Lin [29], Hsu *et al.* [86]). A hyperplane is learned from training instances x_i with class affiliation y_i.

$$\{(x_i, y_i) \mid i = 1, ..., m; y_i \in \{-1, 1\}\} \tag{3.7}$$

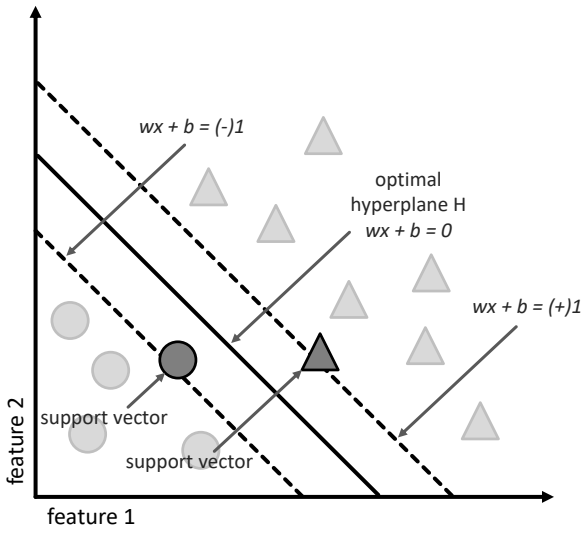


Figure 3.4: Hyperplane for linearly separable classes. Position and alignment of the plane is determined by support vectors.

In case of a multi-class classification problem with k classes $\frac{k(k-1)}{2}$ classifiers with respective hyperplanes for two classes each are learned. A hyperplane H is defined by a

normal vector w and a translation by bias b.

$$H = \{x \mid \langle w, x \rangle + b = 0\} \tag{3.8}$$

To find the optimal values for parameters w and b the following conditions need to be met:

$$wx_i + b \geq 1, y_i = (+)1 \tag{3.9a}$$

$$wx_i + b \leq 1, y_i = (-)1 \tag{3.9b}$$

Which can be subsumed under:

$$y_i(wx_i + b) \geq 1 \tag{3.9c}$$

For linearly separable classes, this can be achieved by solving the optimization problem whilst meeting the condition given above:

$$min_{w,b} \frac{1}{2} \|w\|_2^2 \tag{3.10}$$

In most cases however, classes cannot by separated linearly and we need to allow that made conditions can be broken - but to the most minimal extent possible. To this end, the slack variable ξ and regulation parameter C are introduced in order to keep track and regulate the degree of violation. The optimization problem then looks as follows:

$$min_{w,b,\xi} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i \tag{3.11}$$

The resolution of this optimization problem[2] determines optimal values for vector w and bias b which can now be used in the classification formula:

$$y_i = sign(\langle w, x_i \rangle + b) \tag{3.12}$$

Hence the algorithm uses a linear function to classify data instances, it may not be optimal if the classification problem is not linearly separable. The solution to this problem is the transformation of data into a higher dimensional space with the help of a Kernel function. In the higher dimensional space the data can be expected to be better separable. Possible Kernel functions include:

- The linear Kernel function: $K(x, y) = \langle x, y \rangle$

[2]The way to solve this optimization problem is out of scope at this point.

- The polynomial Kernel function: $K(x, y) = \langle x, y \rangle^d$
- The radial base function (RBF): $K(x, y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)^d$

SVMs are able to model complex and non-linear decision boundaries. It is a popular classification scheme in practical applications as it can be tuned to the given classification problem (Kernel functions) and therefore often delivers high recognition accuracy. The algorithm is however very time and memory consuming.

The SVM approach can also be applied to regression problems (support vector regression - SVR). Hereby, the regression function is given by the minimum of the same functional describing the optimal hyperplane to distinguish discrete classes (Equation 3.10) with the introduction of suitable loss functions.

3.2.3 Bayesian Networks

Bayesian networks (BN) are popular representants of probabilistic graphical models (Ben-Gal [11]). Within a directed acyclic graph, nodes $X_1..., X_n$ reflect random variables that are connected with edges that describe direct conditional dependencies. A directed edge from node X_i to node X_j defines a statistical dependency between the two observations - X_j depends on the value of X_i. The established naming convention denotes X_i as the parent node and X_j as the child node. The value of a node X_i is independent from non-ancestors and only influenced by the given set of parents pa_i. If a variable node has no parents it is called *unconditional*, otherwise *conditional*. The resulting acyclic graph as a whole consequently describes a joint probability distribution over included variables.

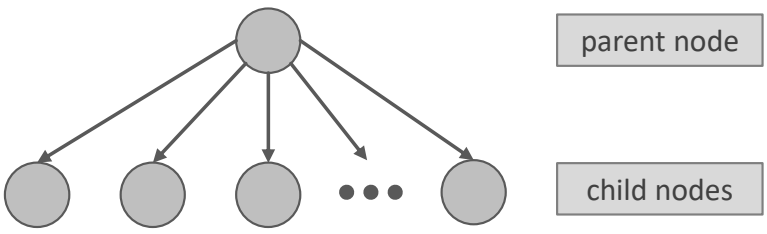


Figure 3.5: Graphical model of a Naive Bayes classifier as simple Bayesian network. Available features are modelled as conditional child nodes of a common, unconditional parent node. I.e. this means that that each feature node X_i has an equal set of parents pa_i.

Like the Naive Bayes classification model (Section 3.2.1), NBs rely on the Bayes Theorem for inferring the joint probability distribution. In fact, a Naive Bayes classifier can be interpreted as the most simple BN which interprets available features as child nodes with a common, unconditional parent node (Figure 3.5). The full joint probability distribution of a BN is given by the chain rule:

$$P\left(X_1, ..., X_n\right) = \prod_{i=1}^{n} P\left(x_i | pa_i\right) \tag{3.13}$$

Hereby, some variables within the BN are given as *observed*. These are called evidence nodes. Nodes with unknown variable values are latent nodes and their values can be inferred from connected evidence node values, the conditional dependencies specified within the respective edges and application of the Bayes Theorem. Consequently, a BN can be interpreted as automatic appliance of the Bayes Theorem to complex problems. Given the structure of a BN (typically defined by expert knowledge), its parameters can be learned from data by finding parameter values that maximize the the (log)likelihood (Friedman [67]) with respect to observed frequencies found within the training data.

A further development of BNs is the dynamic Bayesian network (DBN) (Dagum *et al.* [35]). Here, nodes of the network relate to variables in prior time steps. This way way the node values are not only calculated from current observations but also from past impressions. If this relation is one step to the past (the most commonly used variant), we refer to the model as a two-timeslice BN (2TBN). The addition of further timeslices and relations is of course possible. By transferring this capability to model temporal relations between made observations to the field of multi-modal affect recognition, we have a well suited algorithm to handle affective events in an event-driven fusion approach and will therefore see the respective implementation in comparison to custom approaches in Chapter 8.

3.2.4 Artificial Neural Networks

Artificial neural networks (ANN) were initially designed to simulate the human brain's learning processes as machine learning scheme (Rosenblatt [141]). They describe a network of nodes (neurons), linked by weighted connections (synapses). A multilayer perceptron (MLP) (Rumelhart *et al.* [143]) is a feedforward artificial neural network with multiple hidden layers that connect the input layer to the output layer - the classification result (Figure 3.6). It is generally well suited to solve regression problems, when designed for multi-class discrete classification task, the output layer contains nodes equal to the

number of classes (typically with a softmax activation function, that norms the sum of values of neurons in the output layer between zero and one) whose values can be interpreted as class probabilities.

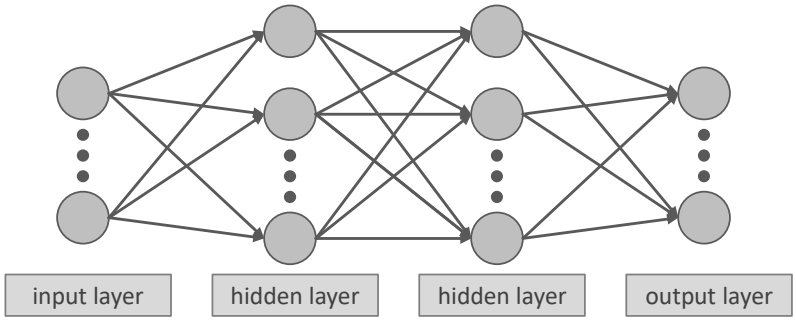


Figure 3.6: Multilayer Perceptron (MLP) with input layer, several hidden layers and an output layer containing the classification result.

Given a hidden layer z_h with $h = 1, ..., s$ neurons connected to input units x_j with $j = 1, ..., t$ and output units y_i with $i = 1, ..., k$ we define weights w_{hj} for edges from input units and weights v_{ih} for edges to output units respectively. Inside a unit z_h of the hidden layer an activation function f is applied to the sum of weighted inbound edges:

$$z_h = f(w_h^T x) \tag{3.14a}$$

with

$$w_h^T x := \sum_{j=1}^{t} w_{hj} x_j \tag{3.14b}$$

Possible activation functions f include:

- Rectifier Linear Unit function: $z_h = max(0, w_h^T x)$
- Tanh function: $zh = tanh(w_h^T x)$
- Sigmoid function: $zh = (1 + e^{-w_h^T x})^{-1}$

During the learning phase, input values of training instances are used to determine all weights within the network via backpropagation.The weights are iteratively tuned until they reliably generate the desired output. Deep neural networks (DNN) that contain several hidden layers have recently become the gold standard for a various of applications, such as speech recognition and image classification.

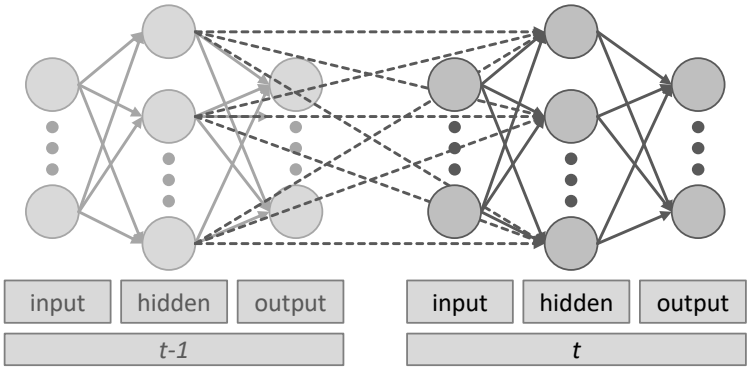


Figure 3.7: Recurrent neural network with input layer, one hidden layer and an output layer containing the classification result. The hidden layer at time *t* has self-connections to the remembered hidden layer of *t-1*.

As with dynamic Bayesian networks, ANNs and especially DNNs can be enhanced with memory capabilities: Recurrent neural networks (RNN) feature cyclic connections. They receive input not only from the input layer (current data), but also from hidden nodes that *remember* previous time steps (Figure 3.7). Although RNNs can handle temporal alignments with this technique, there is a limitation to the range of temporal information that the recurrent network can access (Hochreiter *et al.* [84]). This problem is caused by the so-called *vanishing gradient problem*, which describes the phenomenon of influence of input from hidden layers decaying exponentially.

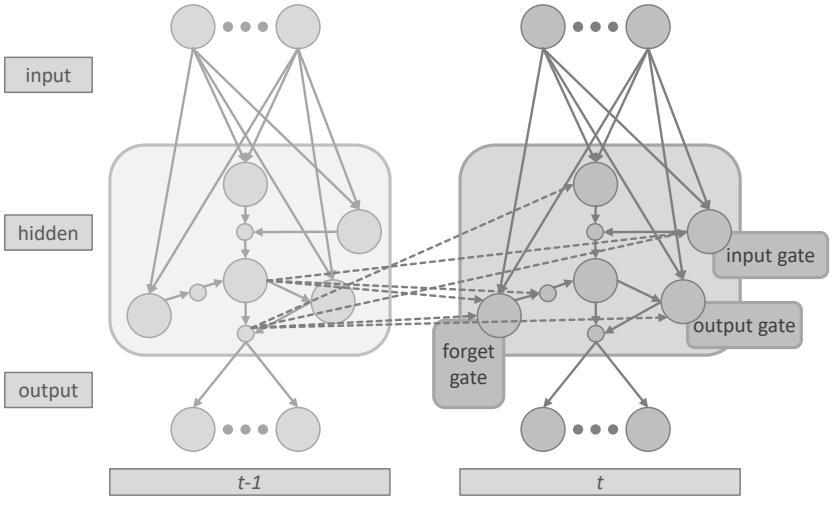


Figure 3.8: LSTM network with one memory block, including the input, output, and forget gate.

Long short-term memory neural networks (LSTM) are devised to better find and exploit temporal context by using memory blocks (Hochreiter and Schmidhuber [83]), that

consist of several recurrently connected subnets: Each memory block contains one or more recurrently connected memory cells and three gate units, the input, output, and forget gates, which control the information flow inside the memory block and are meant to solve the *vanishing gradient problem* (Wöllmer *et al.* [179]). LSTM based models have shown remarkable success in paralinguistic tasks (Brückner and Schuller [21], Wöllmer *et al.* [179]). A possible advancement of LSTM networks are bi-directional LSTM neural networks (BLSTM). These neural networks can access and utilize past and future context and can therefore be expected to yield good recognition results. In multi-modal affect recognition they are able to handle temporal dependencies between modalities and are therefore a comparable approach to event-driven fusion (Chapter 8).

Artificial neural networks are in general able to handle very big input vectors - up to the point where raw data streams instead of descriptive features are fed into the classification system. Irrelevant features tend to be ignored and non-linear decision boundaries are modelled by hidden layers. These advantages make ANNs a very popular classification scheme. Correct network architecture is however a difficult task, as there exist only a few rules of thumb on how to determine parameters such as the number of hidden layers and respective neurons. Calculations are heavily dependent on the weights between neurons and their calculation is hard to reproduce - the term *black box* is often used in this context. The high complexity also leads to high computational costs and time consuming training phases.

3.2.5 Evaluation Methods

In order to judge the quality of a classification system, the classifier is first trained with a set of training samples and is afterwards validated against unseen evaluation-samples. A single iteration is hereby often insufficient. Reasons may be the inadequate size of training and evaluation data, data may contain outliers that sophisticate the result or classification models may depend on random parameters. This leads to the practical approach of conducting several iterations and forming mean values as final and more robust results (Alpaydin and Linke [2]). Of course training and evaluation data needs to be diverse in each iteration and several approaches to generate the needed sub-datasets are in practical use. Bootstrapping (Efron and Tibshirani [53]) and cross-validation (Kohavi *et al.* [101]) are among the most popular ones.

Data Resampling

Bootstrapping is a very simple algorithm that relies on random sampling with replacement. Hereby single instances are drawn randomly from a dataset of size n. Remaining data instances form the training samples. The number of iterations is optional and relies on the complexity of the dataset. The bootstrapped samples may overlap, but the algorithm is nevertheless often used for statistical analyses. The chance to not draw a sample in each iteration is $1 - \frac{1}{n}$. The chance to not draw it after n iterations is:

$$\left(1 - \frac{1}{n}\right)^n \approx exp^{-1} \approx 0.368 \tag{3.15}$$

This means that about one third of samples will not be used for training.

In k-fold cross-validation dataset X is randomly split into k equal parts $X_1, ..., X_k$. In each iteration $k - 1$ parts are used as training data and evaluated against the remaining one. The most extreme form of this validation method is the leave-one-sample-out strategy: For n samples in a dataset, $n - 1$ samples are used in each of n iterations to train the classification system and validation is carried out against the remaining sample. It resembles bootstrapping with fixed n iterations without replacement. This method of course yields often the post possible numerical results but generalisation potential is questionable as often many very similar samples to the tested one are contained in the training set. Given a dataset that comprises data instances of several users, the leave-one-user-out cross validation method may be advisable: For m users in the dataset, samples of $m - 1$ users are used to evaluate samples of user m with m iterations. This approach is very realistic and generalisable as it simulates a user-independent system that is in each iteration only trained with samples from users that are not included in the evaluation set.

Quality Measures - Discrete Classification

The best way to explain quality measures for classification systems is to start with binary classification tasks. In binary classification unseen samples are to be classified into one of exactly two non-overlapping classes (C_1, C_2) (Sokolova and Lapalme [155]). Let us for the following equations denote C_1 as $positive$ and C_2 as $negative$. After evaluating the classification model, we obtain a confusion matrix M:

$$M : \begin{array}{c|c|c} positive & \text{true positive } (tp) & \text{false negative } (fn) \\ \hline negative & \text{false positive } (fp) & \text{true negative } (tn) \end{array} \tag{3.16}$$

The correctly classified members of class C_1 are called true positive (tp), instances of class C_1 incorrectly classified as C_2 are called false negative (fn). Data instances correctly classified as C_2 are called true negative (tn) and members of C_2 incorrectly classified as C_1 are called false positive (fp). Given these categorizations we can now calculate the standard quality measures Accuracy Acc Precision P, Recall R and F-Score F:

Accuracy: Overall effectiveness of a classifier.

$$Acc = \frac{tp + tn}{tp + fn + fp + tn} \tag{3.17a}$$

Precision: Class agreement of the data labels with the positive labels given by the classification system.

$$P = \frac{tp}{tp + fp} \tag{3.17b}$$

Recall: Effectiveness of the classification system to identify positive labels.

$$R = \frac{tp}{tp + fn} \tag{3.17c}$$

F-Score: Relation between data's positive labels and those given by the classification system - also called the harmonic mean of precision and recall.

$$F = \frac{2PR}{P + R} \tag{3.17d}$$

For a multi-class classification problem with classes $C_1, ..., C_k$, we derive the weighted and unweighted average accuracy (WA and UA), which are applied as quality criterias in well known machine learning challenges (Schuller *et al.* [145–149]). For the weighted average accuracy we simply sum up all correctly classified data instances across classes and divide them by sum of all data instances.

$$WA = \frac{\sum_{i=1}^{k} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{k} \tag{3.18a}$$

The unweighted average accuracy is obtained by summing up the precision value for each class and dividing them by the number of classes.

$$UA = \frac{\sum_{i=1}^{k} \frac{tp_i}{tp_i + fn_i}}{k} \tag{3.18b}$$

WA hereby favours bigger classes with more data instances prominent in the dataset, while UA treats every class equal. In case of evenly distributed classes the results are identical.

Quality Measures - Regression

The so far mentioned measures deal with evaluation of discrete classification tasks. When confronted with a regression problem that needs to be evaluated against a continuous annotation, a set of other approaches to determine the prediction quality can be applied. The easiest measure is the mean squared error (MSE) or root mean squared error (RMSE) respectively. These risk functions describe the difference between continuous predictions and the target values and are computed as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.19a}$$

and

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{3.19b}$$

with Y_i as the predicted value at entry i and $\hat{Y}_i$ as the expected value at entry i. These measures are however very sensitive to outliers and give only information about the numerical error. They do not give insights in the shape of the prediction and the annotation. A common evaluation criterion that involves closer description of the course of the prediction is the Pearson correlation coefficient (PCC).

$$PCC = \frac{n\sum_{i=1}^{n} Y_i\hat{Y}_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} \hat{Y}_i}{\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}\sqrt{n\sum_{i=1}^{n} \hat{Y}_i^2 - (\sum_{i=1}^{n} \hat{Y}_i)^2}} \tag{3.20}$$

The PCC measures the linear relationship between two sets of variables. It can range from -1 to 1 where the latter result describes the prediction and annotation following perfectly the same trend, the first a complete countering trend.

3.3 Ensemble Theory

In a single expert system, the categorization output given by the chosen classification scheme may of course be plain wrong. Unseen data of course varies from the limited training set and a single decision boundary will unlikely cover all possibilities. The idea of an ensemble is to combine multiple decision approaches in order to obtain a more informed final result. Figures 3.9 and 3.10 show this strategy: Different decision boundaries or regression functions are applied to the same respective uncategorised samples.

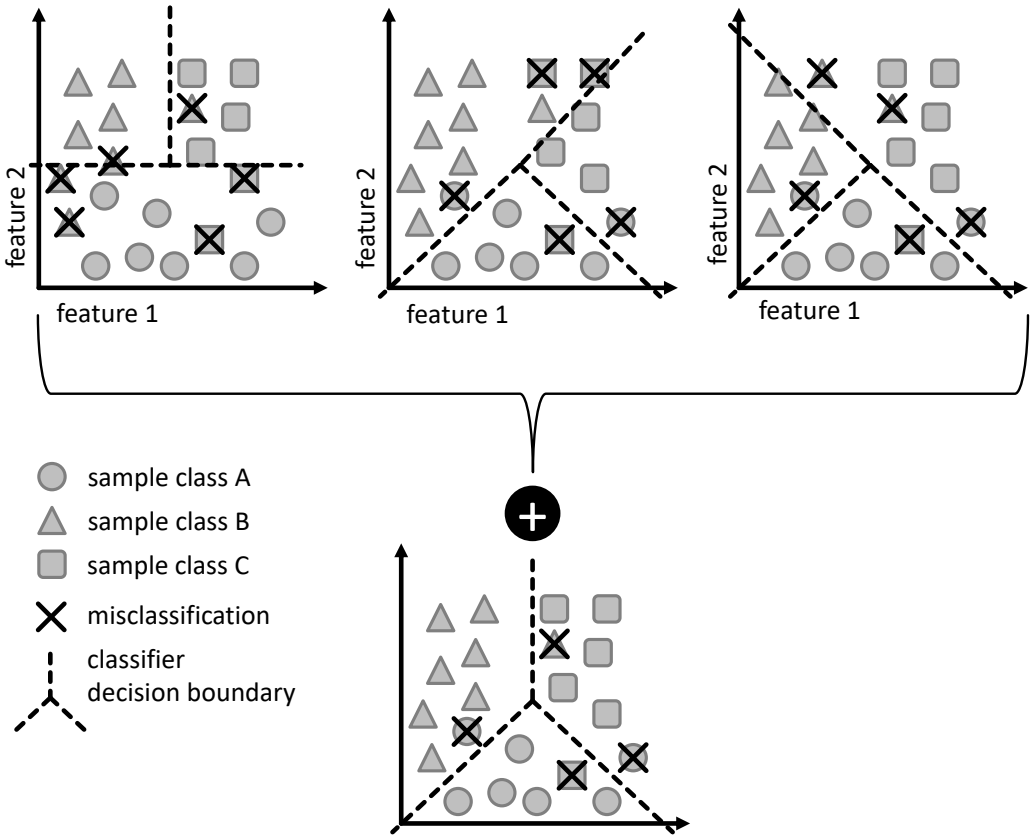


Figure 3.9: Simplified schematic of classifier fusion. Combining decision boundaries of several classifiers can lead to a refinement of recognition accuracy. This can already be achieved by splitting and/or re-sampling of a single data source.

As a result the misclassification rate and regression error is reduced in comparison to the single models.

A fusion system that is aiming to exploit the assumptions of the given ensemble theory consequently consists of a set of different classification models. Given very similar training samples or features for every ensemble member the classification outputs would totally resemble each other and the fusion step would become obsolete. So we have to think of ways to guarantee diverse classification models within the ensemble.

3.3.1 Creating Ensembles

Many classification models offer training parameters that can be set up by the user. Utilising various parameters possibly results in diverse classifiers though employing the same classification scheme. Of course the assignment of completely different classification schemes to the same training data can eventually lead to the desired effect. A more

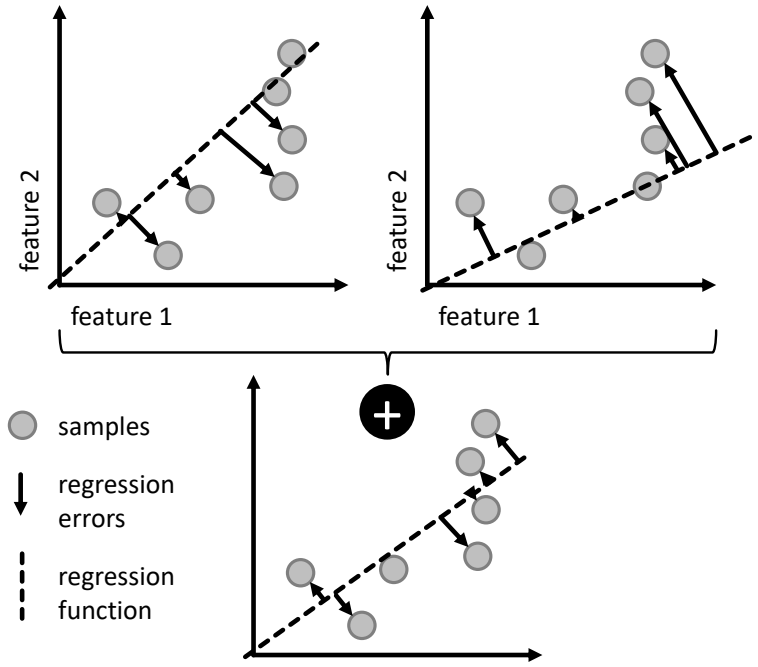


Figure 3.10: Simplified schematic of regression fusion. Combination of multiple regression models can as well lead to better assessments.

common option to achieve the goal of diverse ensemble members is to use different datasets for the classifiers. These can be randomly drawn from an originate dataset by re-sampling techniques. Polikar [135] suggests bootstrapping as the most common one.

In the case of multi-modal affect recognition, the problem of diversity is merely a factor. Where classical ensemble-based machine learning is assuming the reshaping of a single data source for creating diverse ensemble members, we can in most cases start from the premise that observed affective channels deliver differing training data. This of course expects e.g. one classifier per modality. If several recognisers within a fusion system are working on a single source of features, the classical approaches become worthwhile again.

3.3.2 Possible Benefits

Ensemble-based systems can offer some significant advantages over the use of a single-expert system. They are inherent to the ensemble approach and apply whether working with a re-sampled single source of of information or when fusing multi-modal information from different modalities. We have listed the most important ones in (Wagner *et al.* [168]):

Efficiency

In many applications a vast amount of data is gathered and computational efficiency can greatly suffer from training and evaluation of a single classifier with huge datasets. Partitioning of data, training of independent classifiers with different subsets and combination for a final decision often proves to be more practical, time-saving and yields at least competitive results in most cases. Several smaller classifiers may not save training time when using a classification model of linear time complexity (e.g. Naive Bayes - $O(n)$), as the training will consume as much time as for an overarching classifier. But as time complexity rises (e.g. Support vector machines classification scheme (SVM) - $O(n^2)$ or worse) this behaviour changes in favour of small classifiers, using only a section of the original, high dimensional feature vector or training data. Given the contrary case that too little data is available, various re-sampling techniques can be used to form overlapping sub-samples of the dataset. Each of the resulting sub-sets can be applied for training of classifiers, which then are capable of decision making via combination.

Divide and Conquer

Another classification problem can arise if the underlying dataset and the corresponding feature distribution is too complex for a sole classifier to learn. Classification accuracy seriously suffers if the needed decision boundary cannot be found by the used classification model. This undesirable phenomenon can be counteracted by an appropriate set of classifiers. Using a divide-and-conquer approach, the feature space is divided into several (perhaps overlapping) distributions that are easier to learn. Each of these partitions is then handled by one classifier. Adapted combination of the gained classifiers and their simplified decision boundaries adequately simulates the original complex boundary.

Field Performance and Generalisation

The training and testing of classification models typically takes place on data gained from some kind of laboratory environment. Statements about generalized classification performance experienced in field testing - whenever previously unknown samples appear - are difficult to estimate. The risk of performing below average in the field is much higher for a single classifier than for an assemblage of classifier models. Some by chance poorly trained classifiers within a set are a much less of a menace than a single classifier performing poorly.

Handling Unavailable Input

Concerned with systems for multi-modal affect recognition which in practice require the fusion of data from various sensory devices, the case of temporal failures in modalities due to problems, such as hardware breakdowns, operating errors or tracking problems have to be expected. Fusion algorithms can be implemented in a way that they resist unavailability of several input streams. If for example each classifier involved in decision making each represents the observations of an associated sensory device, the absence of a single contribution to the final decision is unlikely to result in a drastic quality fall-off for overall classification accuracy - especially if the sensory malfunction is recognized and the corresponding classifier's (most likely counter-productive) contribution is accordingly rated. Exact suggestions on how to implement correct missing data handling within fusion algorithms will be discussed in Chapter 9.

Chapter 4

Fusion Strategies in Multi-modal Affect Recognition

Uni-modal affect recognition systems rely on a single modality such as paralinguistic features for detecting emotional states. This approach may be too one-sided in order to cope with certain requirements that emerge when trying to detect human emotions. Many emotions are not shown exclusively in a single modality, but rather in a sometimes complex interaction between multiple expressive channels. An emotion like fear may be hard to detect in the voice, but is more expressively displayed in the face. While experiencing sadness, some people may restrain facial reactions to the felt affect, but it can be recognized in the vocal tone. A recognition approach relying on a single modality will in some circumstances only catch a share of the available information. But even if all needed information is theoretically included in a single source, the system becomes useless if the given source is not available. Facial analysis can suffer greatly when the user is not facing directly the camera, vocal investigations are not possible whenever the user remains silent. These problems can be solved (or at least drastically relieved) by applying multiple modalities in an affect recognition system. The scope of available affective information increases, interaction between expressive modalities becomes accessible and temporarily featureless modalities can be substituted by observations from active channels. Fusion strategies meant to achieve these benefits and are commonly used in multi-modal emotion recognition studies will be introduced in Section 4.1. Following this overview we will have a closer look on the appliance and and performance of the differing fusion strategies throughout a survey of recent studies (Section 4.2) and try to extract lessons for our own fusion experiments.

4.1 Levels of Appliance

In order to build an ensemble system for multi-modal affect recognition, a vast amount of eligible fusion strategies to combine the affective channels presented in Chapter 2.2 come into consideration and we will discuss them in detail in the following sections. For the sake of clear arrangement, possible methods can be differentiated by the levels on which they are executed. In a most recent survey D'Mello and Kory [49] group available strategies into fusion at data, feature, decision, score, hybrid and model level. Hereby, naming conventions are not clearly distinguished across studies, fusion at certain levels can be further sub-divided and very sophisticated approaches are sometimes not clearly related to just a single group. In some surveys we find another, less fine grained division of fusion strategies: *early* and *late fusion*. The dividing argument is whether the fusion of information is carried out before or after the first classification step.

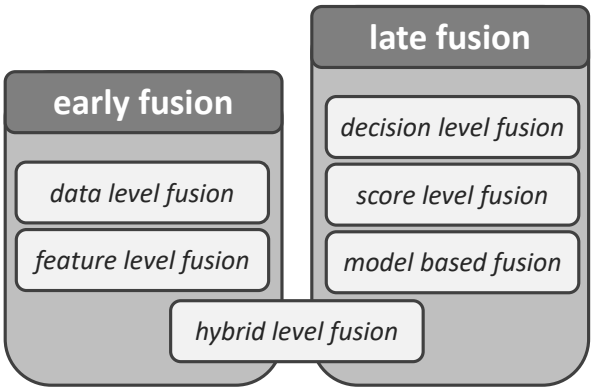


Figure 4.1: Subsuming presented fusion levels under the often used terms *early* and *late fusion*.

Following this convention, we can subsume the levels of appliance presented in this section as seen in Figure 4.1. Whichever naming convention or division of possible approaches is preferred, the following sections offer a good indication on possibilities how to treat the challenge of integrating information from multiple modalities into an affect recognition system.

Nojavanasghari *et al.* [128] present an exemplary comparison of early and late fusion for persuasiveness prediction via deep multi-modal fusion of facial, vocal and textual features in which we can clearly show and compare the differing naming conventions. Nojavanasghari *et al.* use two implementations of multi-modal neural networks for their classification task. In the *early fusion* approach calculated features are merged before they are fed into a deep network topology (Figure 4.2). This approach is de facto a feature fusion approach (Section 4.1.2) with a sophisticated classification model.

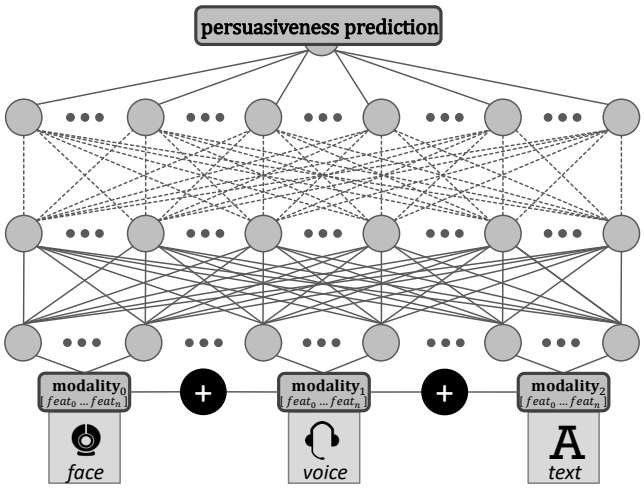


Figure 4.2: In the *early fusion* approach calculated features are merged before they are fed into a deep network topology, which corresponds to a feature fusion scheme.

Two versions of *late fusion* systems are presented in the following explanations (Nojavanasghari *et al.* [128]):

In both cases features of the three individual modalities are fed into separate neural networks. In a first version, probabilistic outputs of modality networks are combined by averaging given scores (Figure 4.3). In a more detailed description, this scheme corresponds to a decision level approach (Section 4.1.3) with an algebraic combination rule 4.1.3).

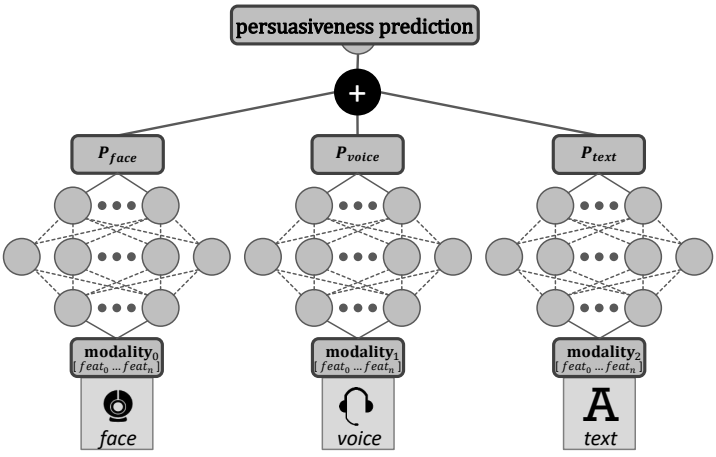


Figure 4.3: In the first version of a *late fusion* approach calculated features of the three individual modalities are fed into separate neural networks. Probabilistic results are averaged in a subsequent fusion step, which corresponds to a decision level approach with an algebraic combination rule.

The second version of late fusion uses an additional deep network topology with calcu-

lated probabilities and their complementary scores as input (Figure 4.4). In other terms this scheme corresponds to a score level approach (Section 4.1.4) with a deep neural network as meta classification scheme.

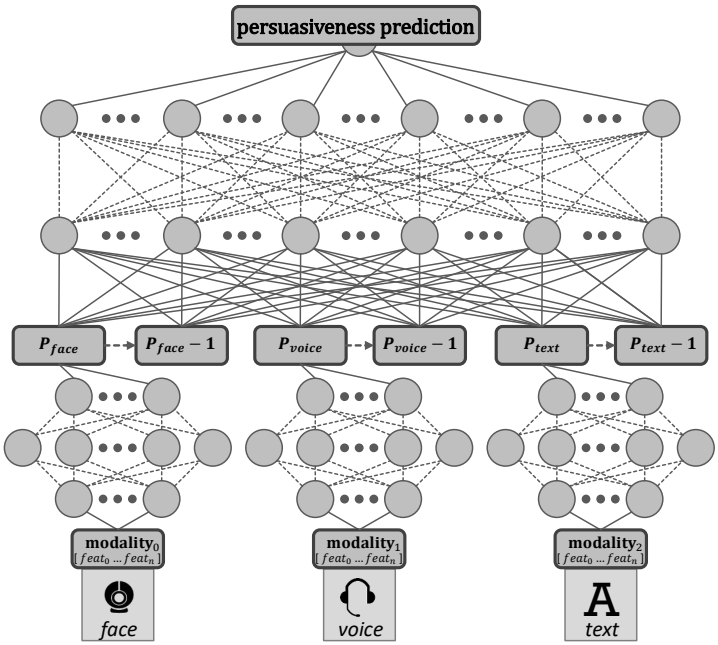


Figure 4.4: The second version of a *late fusion* system also uses three individual modality networks. Calculated probabilities and their complementary scores afterwards serve as input for a network topology that applies the actual fusion step. This strategy corresponds to score level fusion with a deep neural network as meta classifier.

In the evaluation presented within the described study (Nojavanasghari *et al.* [128]), the accuracy of the fusion system rises with complexity and elaborateness. In the case of the study this result serves the discussion well, emphasizing the benefits of a deep neural network as automatic fusion component on top of modality specific classifiers. However, we will see in upcoming chapters and evaluations that such findings are hard to generalize.

4.1.1 Data Level Fusion

Fusion on data level is a traditional approach to combine information from multiple sources. Improvements in processing hardware during the 90's enabled the parallel treatment of multiple sensors with the goal to achieve improved accuracies than could be achieved by the use of a single sensor alone (Hall and Llinas [77]). The most prominent examples of this type of data fusion can be found in the field of image processing.

Here, pictures of multiple camera sensors of different sensor types are merged to reduce artefacts and reveal information that can better be recognised based on mutual observation. Examples include the merging of computer tomography and magnetic resonance images in medical applications (Jamee *et al.* [90]) or the detection of concealed weapons based on RGB and infrared images (Xue and Blum [183]) as well as shots taken from different angles and positions. By definition, data level fusion includes all combination techniques that fuse information gained from modalities on the signal level, i.e. before the feature extraction step. Generally, data level fusion is hardly found in emotion recognition studies. In our own experiments we use data level fusion within a sensing architecture meant to detect implicit user reactions (i.e. user arousal) within an immersive mixed reality installation (Omedas *et al.* [129], Wagner *et al.* [171]). Based on wearable sensor devices, physiological signals were captured and the normalized heart rate signal was merged with the normalized phasic component of the galvanic skin response. As both signals feature the characteristic to increase in amplitude and form local peaks in phases of high user arousal, the fused signal can be expected to show these characteristics intensified when arousal is observable in both modalities and still in measurable quantity whenever high arousal is only shown in one of the channels.

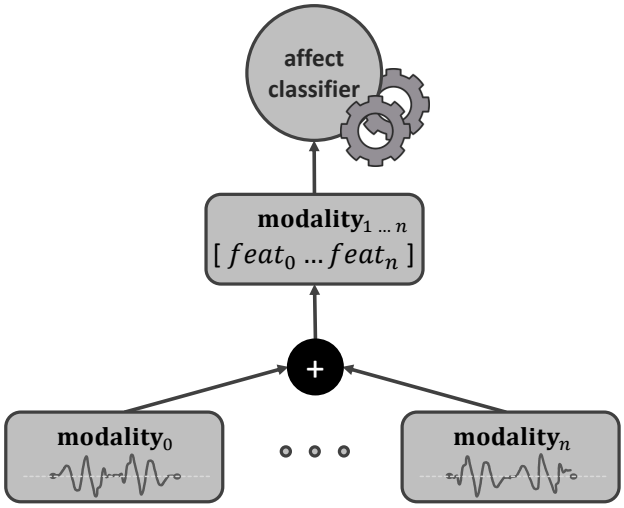


Figure 4.5: Data level fusion schematic. Signals from multiple modalities are combined before the feature extraction step.

4.1.2 Feature Level Fusion

Feature level fusion is a common and straightforward way to fuse all recorded observation-channels. All desired features are merged into a single high dimensional feature set.

One single classifier is then trained for the task of affect classification (Figure 4.6). As there is only one classification model used, this fusion strategy cannot be counted towards ensemble-based techniques. Because the fused data contains a bigger amount of information than single modalities, an increase in classification accuracy can theoretically be expected. These classifiers are often used in emotion recognition studies and tend to yield reliable classification results (Lingenfelser *et al.* [110]). The combined statistical features of several affective channels form a consistent impression of affective hints available in the available modalities within a fixed segment of time. Due to its simplicity, feature level fusion is often used in studies as a multi-modal baseline system to compare more sophisticated fusion approaches to.

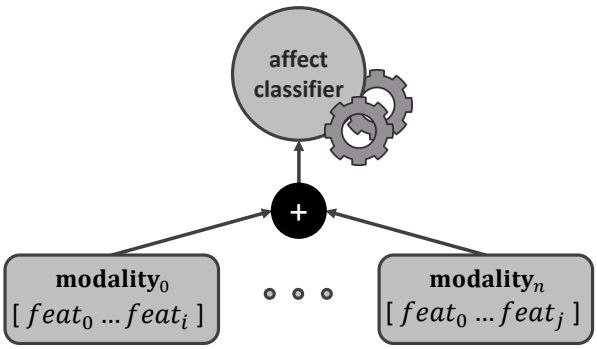


Figure 4.6: Feature level fusion schematic. Information from multiple modalities is combined by concatenating respective feature vectors for classification.

This very accessible approach to multi-modal fusion comes along with a couple of problems: One drawback is the eventually occurring curse of dimensionality on small datasets. If the available data is not ample, the classification results become non-meaningful. It has to be mentioned that a growing feature vector may stress computational resources for training and evaluation of the classification model. Appliance of feature selection techniques may however relieve both problems. In most examinations these obstacles may not be of interest due to a fair availability of time and resources, other ones (e.g. mobile applications) may refuse the feature level approach solely because of these reasons.

Before surveying modern fusion systems (D'Mello and Kory [48, 49]), D'Mello and Kory used this rather simplistic fusion approach merge features from conversational cues, body postures and facial expressions to detect experiences of boredom, engagement, confusion, frustration and delight (D'Mello and Graesser [47]). To enhance the multi-modal approach, they use feature selection on the single affective channels to only include the most informative features in the final fusion step. Castellano *et al.* [28] apply the same modalities to recognize eight basic emotions, such as anger, joy or fear. Feature level

fusion is compared to uni-modal accuracy and a simple fusion mechanism on the decision level - majority voting. In the according evaluation study feature level fusion (Figure 4.6) outperforms uni-modal classification by more than 10% and even decision level fusion by more than 4%, underlining the good capabilities of the most straightforward way to fuse multiple modalities.

4.1.3 Decision Level Fusion

Another very popular fusion method in affect recognition is decision level fusion. The term sums up combination rules for the outputs of several classification models. Contrary to feature level fusion and its reliance on a single classifier that deals with a high dimensional feature vector, decision level fusion focusses on the usage of small classifiers and their combination. The available feature set is divided into subgroups and the partitions are used to form several small classification models (of course these classifiers can also be generated by sub-sampling training data or the usage of different classification models). Outcomes of these slim classifier models are taken into account for the final decision making process. This strategy is very desirable in affect recognition, as often affective channels with very diverse features are observed and to train one model per modality is a natural and convenient to handle the integration of multi-modal information (Figure 4.7). Though there are very elaborate algorithms to combine the outputs of the ensemble classifiers, there are also rather simple combination rules available that allow a straightforward implementation.

Based on the algorithm chosen to fuse the decisions of ensemble members, we can further subdivide this group of fusion techniques. Therefore we need to go into detail and review the combination logics.

Class-Label Combination

The most generic approach to merge the classification output of multiple affective channels is to have each modality decide for one of the desired target classes and subsequently hold a poll for final decision. No statement about certainty of the chosen class is needed, the fusion algorithm only processes the definite class labels. A way to refine this approach is to use previous knowledge about available information sources. If we face a multi-modal affect recognition scenario and know from observation, that one modality delivers more accurate assessments than the other channels, we can chose to to give a higher weight to the respective classifier's vote. In 2012 Koelstra *et al.* presented

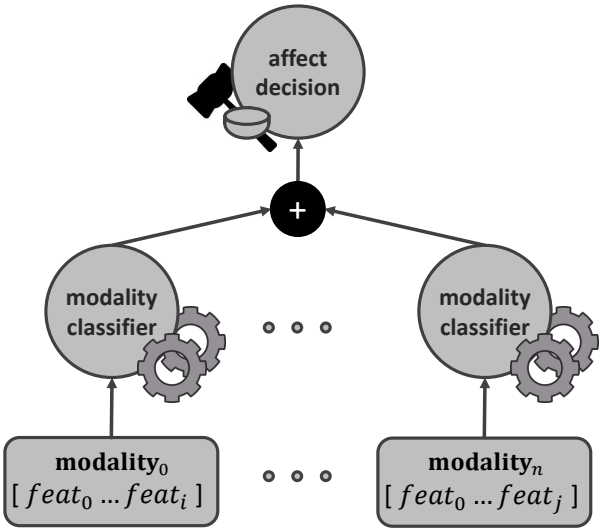


Figure 4.7: Decision level fusion schematic. Different modalities undergo separate classification steps and the fusion is applied based on the resulting decisions.

the very interesting and well known database for emotion analysis with physiological signals (DEAP) (Koelstra *et al.* [100]). Here, emotions are induced by presented music videos and an exhaustive selection of peripheral physiological signals (including galvanic skin response, blood volume pressure, respiration, skin temperature, muscle activity and eye movement) is recorded in combination with power and spectral features from clinical electroencephalography. The work includes a study on fusion strategies for the multiple modalities: Physiological features are merged into one modality and fused with features from electroencephalography and paralinguistic features extracted from the music videos associated with each recording sample. Koelstra *et al.* use feature level fusion as baseline system and emphasize an advantage of decision level fusion, namely the possibility to apply weights to the given modalities. By adopting a grid search mechanism during the training step, they are able to find an optimal weight distribution among their modalities, which is able to improve the performance of the recognition system compared to feature level fusion and unweighted decision level fusion.

Another way to refine the processing of class labels given by ensemble members is to not use a simple voting mechanism, but to keep track of the decisions in so called lookup tables and learn which distribution of votes is most often associated to each true class. The original fusion concept was introduced in 1993 by Huang and Suen [88], but is problematic on datasets of relatively small sample size - which unfortunately is often the case in multi-modal affect recognition studies - because of overfitting of the lookup tables. However, the approach is still subject to improvements (Raudys and Roli [139]), and we present an approach to solve the problem for the small datasets in our first fusion

experiments (Chapter 5).

Algebraic Combination Rules

Algebraic combination rules take more information into consideration than just the final decision of ensemble classifiers. The before mentioned voting mechanisms are replaced by mathematical computations and in contrast to class label combination, assertions about the calculated class probabilities are needed as input. The most simple decision rule given the probability distributions across all classifiers would be to choose the the decision which was made with highest certainty. Classically however, the given probabilities for each class are combined over all classifiers by calculating e.g. the sum, product, mean or median. The final decision is consequently based on comparison of these average values. For emotion recognition purposes, we can hope that this approach gives each modality a chance to contribute to the final assessment to an extent that reflects how clearly the affect was recognizable in each channel. Kanluan *et al.* [92] present a prototypical fusion system that uses weighted linear combination of audiovisual estimations (the most dominant modality combination across present fusion systems (D'Mello and Kory [49])) on decision level and also conclude very exemplary findings concerning dimensional models of emotion: Results showed that arousal related emotions can be best recognized with acoustic features, visual features are better suited for valence estimation. A combination of both on decision level is in consequence desirable.

Specialist Selection

The exploitation of prior insights (e.g. gained from from previous recordings or training data) can be further utilized by finding modalities and corresponding classifiers in a given ensemble that are exceptionally well suited to recognize single classes/emotions within the observed spectrum. This knowledge is used by specialist selection algorithms in order to increase overall accuracy of the fusion system. In Chapter 5.1.3 we introduce an exemplary implementation of this approach (Kim and Lingenfelser [97]) that employs specialist classifiers trying to identify weakly recognized classes before the whole ensemble makes its decision. In the case of emotions we hope to find more subtle depictions of affect that can only be observed in certain modalities, which would otherwise be overwritten by the impressions from other channels.

4.1.4 Score Level Fusion

In score level fusion, the outputs of several ensemble classifiers are not fused by predefined combination rules. Instead their results are used as input for one or more meta classification models, that generate the final ensemble decision. This process is lent from meta-classification and conforming to notations used by Wolpert [180], ensemble classifiers correspond to so-called level-0 base classifiers, the meta classifiers fusing their results equate to level-1 meta generalisers. The fusion process is very related to decision level fusion and often subsumed under the term, but the difference of having an automated classification process for fusion versus following predefined rules makes these approaches worth mentioning separately.

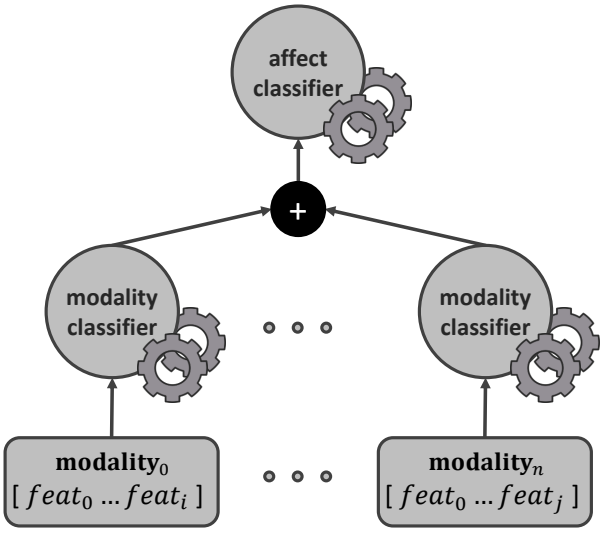


Figure 4.8: Score level fusion schematic. Instead of combining the probabilistic outputs of several classifiers on basis of predefined rules, they are used as input for meta classifiers that produce the final fusion result.

Classical implementations of score level fusion are stacking and grading: In stacked generalisation - as proposed by Ting and Witten [159] - a level-1 classifier tries to learn the probability distribution among level-0 ensemble classifiers together with the true class that lead to this combination. When asked to classify an unknown sample, the method first collects probability estimates of all ensemble members that consecutively form the basis for the level-1 classifier's final prediction. Another approach to meta-classification is grading (Seewald and Fuernkranz [151]), where the goal of level-1 classifiers is to correct potentially false decisions of level-0 ensemble members. During training every base classifier is complemented by a meta classifier with same training data but a graded label - a boolean value stating correct or incorrect prediction of the ensemble classifier. At

classification time ensemble predictions are fused as every member adds the probability of correctness (generated by its grading classifier) to the final support of the class it predicted. As usual, the class with highest support is chosen as final ensemble decision. The possible advantage of this fusion approach for emotion recognition is, that no deeper understanding of the underlying (and maybe very complex) emotion model is needed as no combination rules have to be chosen. Instead the whole decision process is automated by classification models.

4.1.5 Hybrid Level Fusion

Hybrid level fusion is a wide category subsuming fusion approaches that - in the case of affect recognition - contain classifiers based on features from single modalities as well as ones that contain merged feature sets from multiple modalities within their ensembles. Therefore these fusion systems can be described as a mixture of feature and decision level fusion (Figure 4.9).

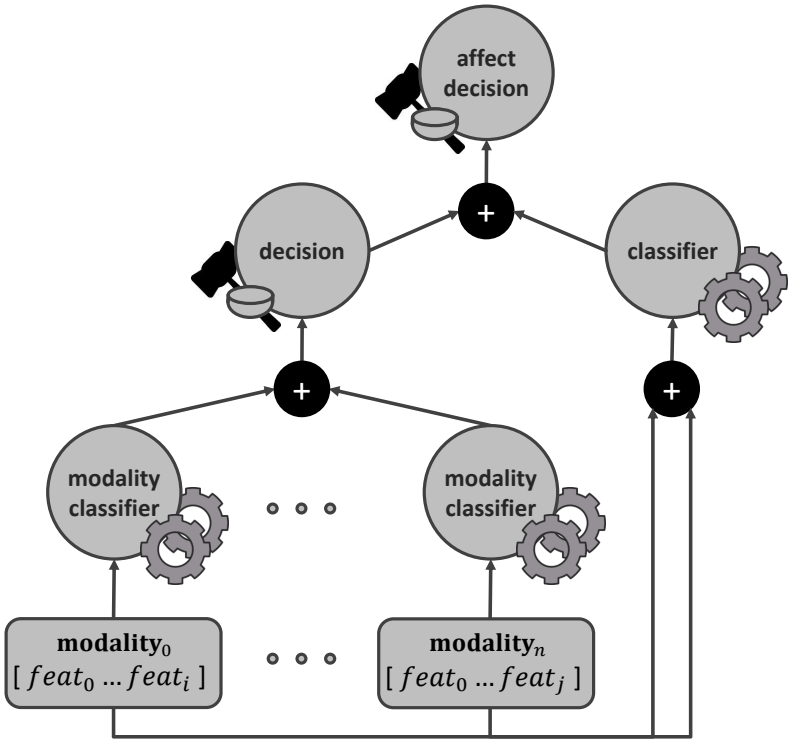


Figure 4.9: Hybrid level fusion schematic. A mixture of feature and decision level fusion in which the fusion system contains classifiers based on features from single modalities as well as ones that contain merged feature sets from multiple modalities.

Kim *et al.* [98] use hybrid level fusion in a bi-modal emotion recognition task where

features from various physiological channels (electromyography, electrocardiography, skin conductivity and respiration activity) are used in combination with paralinguistic features to classify the four emotion quadrants of the valence-arousal space (Chapter 2.1). They extend majority voting on decision level by adding a classification model-based on all available features to the ensemble which otherwise contains only classifiers trained on the single affective channels. This simple addition leads to accuracy improvements in most parts of the included user dependent evaluation. In [31] Chetty *et al.* describe a multilevel fusion approach for audiovisual emotion recognition in which features from visual data are calculated from differing areas of the face (forehead, mouth, eye, left and right cheek). Each face area is given a classifier and a merged visual feature model is used to train an additional face classifier. These, together with the prosody classifier form the audiovisual hybrid fusion system which is evaluated to perform significantly better than feature and decision level fusion on the DaFEx emotion corpus (Battocchi *et al.* [8]).

4.1.6 Model-based Fusion

The term model-based fusion is used on a very broad scale. In [49] D'Mello and Kory define all fusion methods that aim to model temporal dependencies between used modalities as model-based fusion approaches (Figure 4.10). If we imagine the audiovisual display of a positive emotion, we expect a certain tone of voice or even laughs or giggles accompanied by signs of positiveness like smiles or grinning in the face. Combining the information from modalities should shape a clear fusion result. The problem, however, is, that these affective cues are not guaranteed to be present within the same analysis window, occurrences can be time shifted across modalities.

As an example, the fusion system can recognize a positive tone in the voice while no matching facial expression is shown within the timeslice. If the fusion scheme does not model temporal interaction between channels, the system may be confused by contradicting cues. If however temporal alignments are modelled, the system may be aware that a matching facial expression was very well present within the last windows and include this information in decision making. In order to do so, the fusion scheme needs probabilistic connections to surrounding analysis frames - or in the case of real-time systems, past frames (Figure 4.10).

Dupont and Luettin [52] were among the first to tackle the asynchronous nature of audio and video streams by modelling temporal topologies with multi-stream hidden Markov models for continuous speech recognition (Figure 4.11). Here modalities are processed independently with distinct hidden Markov models until problem-specific anchor points

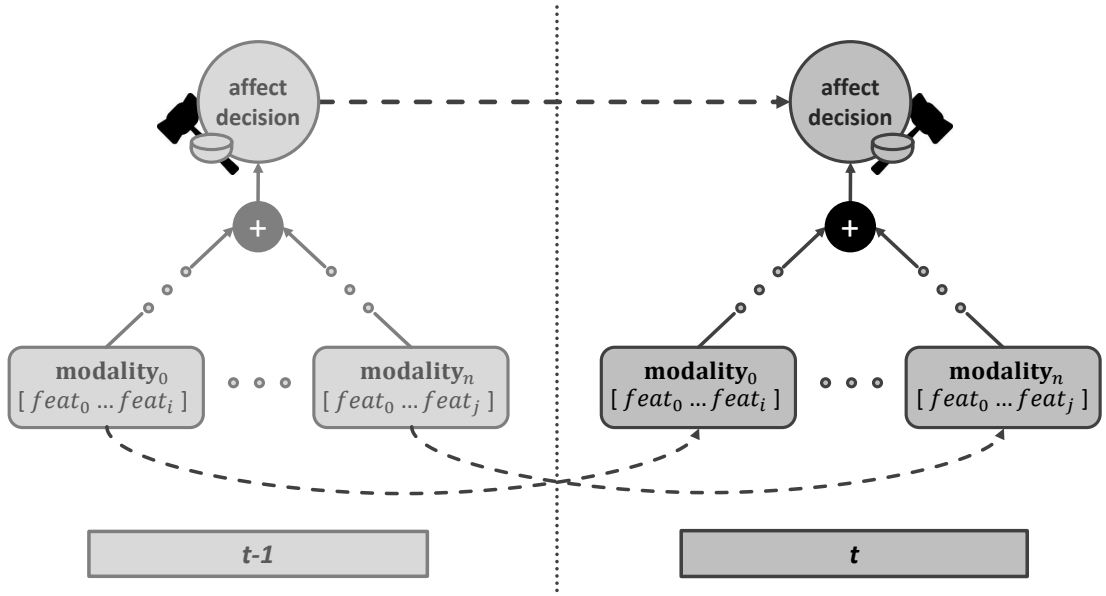


Figure 4.10: Model-based fusion schematic. Temporal dependencies are incorporated into the fusion process by considering past time frames.

(e.g. ends of phonemes, syllables or words) are reached. At these synchronization points the individually accumulated scores over the current temporal segment (i.e. the segment given by the distance to the last synchronization point) are combined. This means that diverse temporal flows within modalities can be modelled, sample rates of affective channels may be asynchronous, but the anchor points and resulting segments dictate synchronicity of decision boundaries.

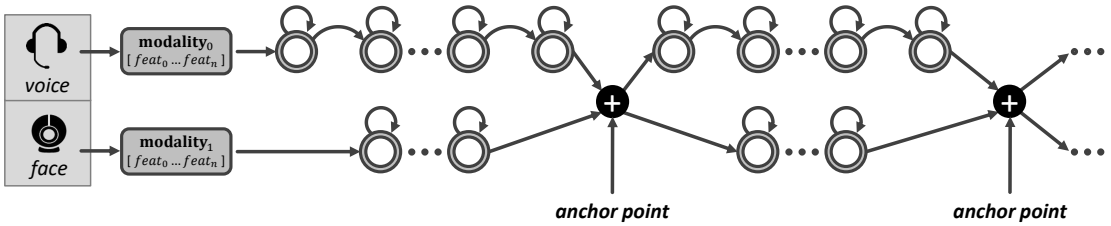


Figure 4.11: To tackle the asynchronous nature of audio and video streams, modalities can be processed independently with distinct hidden Markov models until problem-specific anchor points are reached. These take care of a synchronous beat of decisions.

Another prominent approach to model temporal dependencies between modalities is presented by Glodek *et al.* in [72] and [73]. They propose a Markov fusion network that combines a time series of decisions of independent classifiers. The classifiers are given a reject option if confidence is not sufficient, resulting in a sparse decision vector

with missing input. During an estimation at time t, the algorithm is aware of current classifier decisions and network estimations at time $t-1$ and $t+1$. The Markov assumption is used to reconstruct missing classifier inputs (caused by the reject option) and should smooth outliers and misclassifications. We covered a few earlier examples of temporal, model-based fusion approaches that try to exploit cross-correlations between audiovisual modalities in (Wagner *et al.* [168]): Song et al. [158] propose a tripled hidden Markov model that models correlations between features from upper face, lower face, and prosody. Via relaxing the general requirement of synchronized segmentation for audiovisual streams, Zeng et al. [185] introduce a multi-stream hidden Markov model which enables an optimal combination of multiple streams from audio and video modalities. The maximum entropy and the maximum mutual information criterion are used in order to estimate the correlation between considered streams. Sebe *et al.* [150] model the interdependencies between audio and video channels with dynamic Bayesian networks. Imperfect data is hereby handled via probabilistic inference. In more recent systems, we see that numerous variants of recognition models that try to describe multi-modal correlations are based on recurrent neural networks, e.g. (Fragopanagos and Taylor [64]) and (Caridakis *et al.* [25]). Figure 4.12 shows a current approach to model-based fusion presented by Ringeval *et al.* [140] by modelling the decision of each modality within (bi-directional) long short-term memory recurrent neural networks ((B)LSTM-RNNs). Results of the neural networks are fused with a support vector regression (SVR) model. This individual treatment of each modality allows differing analysis windows and sample rates, the memory capability of recurrent networks (Chapter 3.2.4) includes a temporal component of recognition. Temporal dependencies across modalities, however, seem to be not considered by this fusion model as the instance combining the modalities (SVR) does not feature any memory capabilities.

4.2 Performance of Multi-modal Fusion

Considering the so far discussed advantages of multi-modal fusion in addition to the inherent benefits of ensemble-based recognition approaches (Chapter 3.3.2), we should be expecting an almost guaranteed increase in recognition accuracy whan applying these techniques to an affect recognition system. However, the effect of multi-modality in comparison to a uni-modal system is reported to be sometimes minimal and sometimes even counter-productive: In Chapter 1.1 we introduced the findings by D'Mello and Kory [48], stating the mixed impression gained on the effect of multi-modality in affect recognition systems after reviewing 29 contemporary and relevant studies. In 2015 this

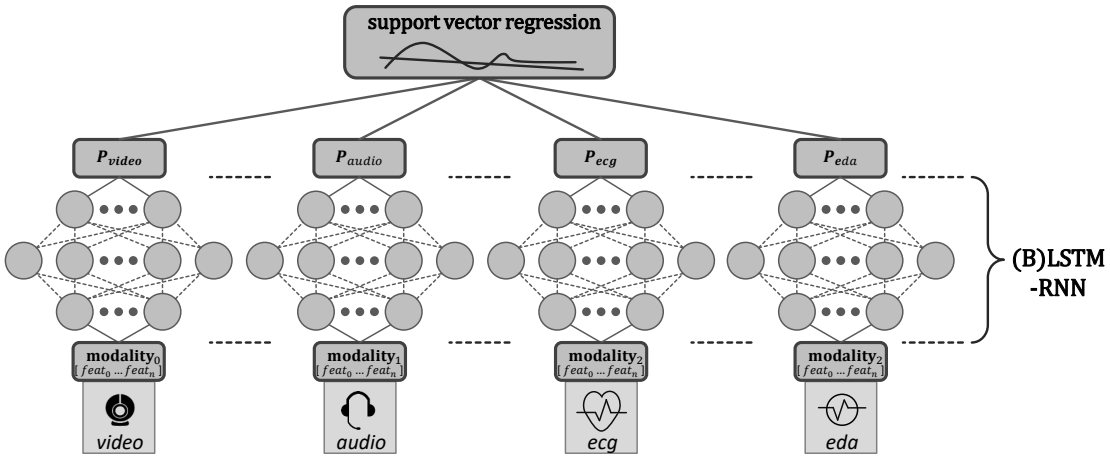


Figure 4.12: Individual treatment of each modality with (B)LSTM-RNNs allows differing analysis windows and sample rates. The memory capability of recurrent networks includes a temporal component of recognition within modalities, but not across affective channels.

meta review was greatly extended to a total of 90 peer-reviewed systems (D'Mello and Kory [49]). These studies are investigated very detailed in terms of used emotion models, modalities, fusion methods, the general performance and especially the multi-modal effect on classification accuracy in comparison to uni-modal classification.

4.2.1 Multi-modal Effect

The intention of considering more than one modality for a given emotion classification problem is that additional sources of affective information enhance recognition performance. In addition to the increase in available information, temporal interaction between modalities can be detected (Section 4.1.6) and exploited for increased accuracy. Additionally the whole affect recognition system should theoretically gain robustness, as modalities that occasionally won't contribute meaningful information can be substituted by other channels.

To present the hopefully positive impact of multi-modality on an affect recognition system in numbers, D'Mello and Kory [49] define the multi-modal effect (MM) as the percent improvement of multi-modal fusion accuracy ($fusion$) over the best of n uni-modal accuracies ($modality_{(1,\dots,n)}$) presented within respective studies.

$$MM = 100 * \frac{fusion - max(modality_{(1,\dots,n)})}{max(modality_{(1,\dots,n)})} \tag{4.1}$$

Given this definition several statements on the 90 peer-reviewed systems can be made: Over 85% of studies report a multi-modal effect greater than 1%. This in turn means that 15% of studies did not gain any significant improvement by including additional affective channels in the recognition process. Some even report a negative multi-modal effect.

The reasons for a low positive or negative multi-modal effect are different from study to study. Unsuitable modality selection for a given classification problem or inappropriate features can cause a negative effect on the whole recognition system. Sometimes one modality is clearly superior to other channels and the more unreliable information from additional sources rather confuses the system than enhancing it. A general trend however is clearly visible, i.e. a tendency to achieve a lower multi-modal effect on natural data than on acted datasets. Acted emotions often are depicted with increased intensity which results in more expressive features. Within the acted samples of a single emotion class there is decreased variability to be found. When acting an emotion, most actors have a stereotypical display in mind and show it with only slight deviations. This homogeneity and absence of outliers eases the calculation of a decision boundary. Another observation is the increased coordination between modalities. A natural unconsciously shown emotion may only be observable in a single channel at a time while an acted emotion may be intentionally portrayed in all affective channels at once. A multi-modal fusion system receiving these acted inputs can work on coherent cues in modalities. On natural data, the fusion process might deal with sparse information.

Overall, the survey reports a modest gain (9.83% mean and 6.60% median improvement) of multi-modal fusion over uni-modal recognition systems. For reasons discussed above, the best multi-modal effects occur in studies that use acted data for evaluation. Furthermore, model-based fusion systems that incorporate the temporal flow of modalities tend to achieve better results than their more simple counterparts.

4.2.2 Implications and Expectations for Upcoming Experiments

From the findings presented in (D'Mello and Kory [49]) we can derive implications for our own experiments we are going to present in the following chapters. First off, the accuracy of multi-modal fusion can be expected to enhance affect recognition results compared to uni-modal classification - but to a moderate extent. The performance of fusion systems is dependent on the quality of the single modalities and will not raise the recognition quality disproportionately.

Discrepancies between the use of datasets depicting acted and those showing natural emotions have by now become obvious. These could have consequences for affect recognition systems that are meant to be applied in real-world applications, where the display of natural emotions does not fit the acted training data. We will therefore investigate the performance of fusion systems on acted and naturalistic data more detailed in Chapter 6.

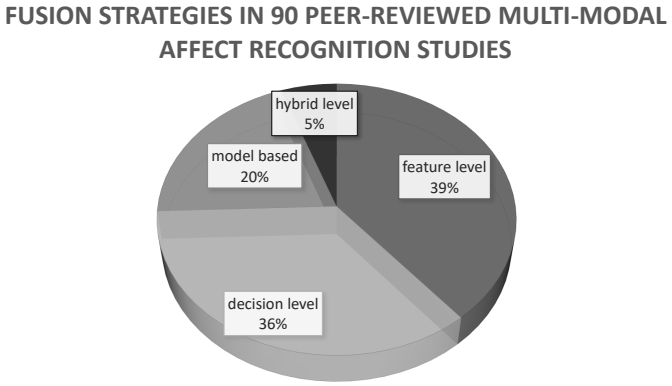


Figure 4.13: Model-based fusion approaches are used in 20% of reported affect recognition studies. Implementation effort required to realise asynchronous fusion algorithms tends to be higher than for synchronous approaches.

In Chapter 7 we are going to present our own approaches to asynchronous and event-driven fusion techniques which count towards the model-based fusion schemes. Consequently findings concerning such fusion systems are of special interest at this point and we will now conduct our own interpretation of the presented studies. The modelling of temporal dependencies has lead to multi-modal affect recognition systems that range within the most promising fusion approaches of the discussed survey. However, if we count reviewed studies applying model-based approaches, we see that these are only applied in 20% of evaluated fusion systems (Figure 4.13). This may be on the one hand explained by the novelty of these systems compared to more traditional fusion schemes. On the other hand, the implementation effort required to realise these asynchronous algorithms is significantly higher than e.g. synchronously concatenating feature vectors for feature level fusion.

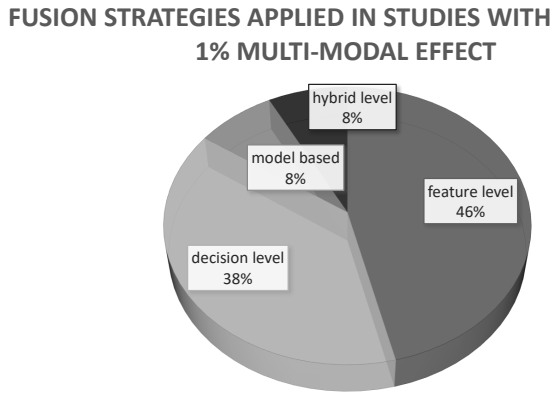


Figure 4.14: Though model-based fusion approaches are used in 20% of reported affect recognition studies, they only occur in 8% of the experiments that report a multi-modal effect below 1%.

If we investigate of the survey data even closer, we find that model-based fusion approaches appear only in 8% of studies reporting a multi-modal effect below 1% (Figure 4.14). This reveal implies that model-based fusion (and therefore the consideration of temporal alignments) deals better with settings that remain a problem for traditional approaches. These include naturalistic depictions of affective states as well as temporary unaligned cues across affective channels. We will present a detailed discussion of possible advantages model-based fusion may offer in these scenarios in Chapter 7.

Chapter 5

First Experiments: Standard, Custom and Emotion-adapted Fusion Strategies

Now that we have established the theoretical basis for emotion modelling and affective channels (Chapter 2), ensemble-based machine learning (Chapter 3) and the resulting strategies for multi-modal fusion in affect recognition (Chapter 4) we can now conduct the first practical experiments. In Section 5.1 we present a custom approach to specialist selection and decision level fusion and compare it to more standard fusion techniques. Section 5.2 will introduce our concept of an emotion-adapted fusion approach. Generally said, an emotion-adapted combination strategy is not universally applicable for any given classification problem, but logically tied to an an underlying emotion model.

5.1 A Custom Approach to Decision Level Fusion

The experiments we are going to carry out in this section do not genuinely fall under the category of multi-modal affect recognition, as target classes do not consist of emotional states and only the vocal modality is available. However, we will herein exploit the possible benefits of ensemble-based decision making as described in Chapter 3.3.2 and apply fusion techniques introduced in Chapter 4.1. More precisely, ensemble members are generated by providing multiple feature sets generated by feature selection, and novel fusion methods developed in order to give special support to under-represented classes are introduced for decision making. Results are compared to standard classification approaches and possible benefits are discussed.

5.1.1 Age and Gender Recognition from Speech

In a contribution (Lingenfelser *et al.* [109]) to the INTERSPEECH 2010 Paralinguistic Challenge (Schuller *et al.* [145]) we explore the capabilities of standard and custom decision level fusion and ensemble-based techniques for classification tasks on the provided aGender[1] corpus: Humans are used to adapt their behaviour in dependence of age and gender of their communication partner. Likewise, a system could involve this contextual information about its users in the decision making process to respond in a more adequate way. Using paralinguistic information conveyed in speech offers a possibility to automatically detect a speakers' age and gender, and has evolved to an own sub-discipline in the field of speech analysis contributing to speaker classification (Müller [123]). Yet, at the time given no standard regarding evaluation procedures and comparability had been established. In response to this deficit, the INTERSPEECH 2010 Paralinguistic Challenge addressed age and gender classification in two sub-challenges. Before, age and gender recognition from speech has often been treated as a combined problem. Müller [123] uses a variety of paralinguistic features including pitch, voice quality, speech rate and pauses to recognise eight classes (four age classes separated by gender). Bocklet *et al.* [14] compare Gaussian mixture models (GMMs) and support vector machines (SVM) on cepstral features only to recognise seven gender/age classes.

5.1.2 Building a Diverse Ensemble from a Single Modality

While common classification approaches focus on creating a single expert classification model which uses a high dimensional feature vector as input, the terms ensemble techniques and fusion sum up an array of methods that rely on the generation of some slim classifiers and the associated combination rules. Neither must these classifiers provide perfect performance on some given problem, nor do their outputs need to resemble each other. It is sometimes even preferable that the chosen classifiers make mistakes, at best on different instances. A basic idea of decision level fusion is to reduce the total error rate of classification by strategically combining the members of the ensemble and their errors. Therefore the single classifiers need to be diverse from one another (Wagner *et al.* [168]). In multi-modal affect recognition this task can easily be achieved by providing each classifier with features from one of the available modalities. But as in this case only one modality (i.e. speech) is available, we need to find ways to produce diverse feature sets within the available channel.

[1] `https://www.phonetik.uni-muenchen.de/Bas/BasaGenderdeu.html`

In a preceeding work by Schuller *et al.* [144] on emotion recognition, bagging and boosting algorithms are applied to create a diverse ensemble of classification models from acoustic and linguistic features extracted from the audio channel. The goal hereby is to avoid data overfitting in situations with a relative small number of training samples compared to a high number of features. Score level fusion (i.e. stacking) is used to combine the outputs of generated ensemble classifiers. Using the proposed techniques the authors could achieve a slight improvement from 70.27% to 71.62% on an eight class problem compared to the best single expert classification system with SVMs. Lee *et al.* [108] propose a hierarchical tree of binary classifiers for affect recognition from audio data. They select the individual feature sets for each tree node from the same global feature set and derive the tree structure from the difficulty of the classification subtasks. This way the individual classifiers are not generated by sub-sampling techniques, but by diverse feature sets for each model. With this framework the authors were able to improve the unweighted accuracy in a five-class problem by 3.37% compared to a single SVM classifier.

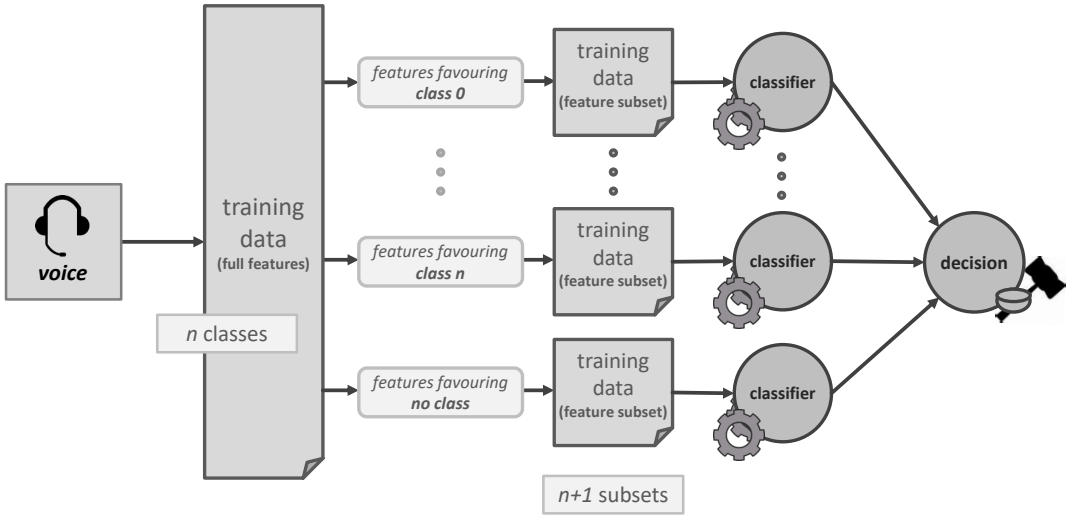


Figure 5.1: The final recognition system is composed of a fusion strategy and the diverse classifier ensemble gained by modified feature selection steps.

In the given case we try to achieve the crucial requirement of diversity by a variation of available features through modified feature selection appliance: The classifiers used in creating the ensemble for all discussed decision making algorithms stem from different feature sets generated by modified feature selection processes. More precisely, for n given classes of a classification problem $n + 1$ feature subsets are generated (Figure 5.1). The first n feature sets are chosen with preferential treatment of the respective n classes. In addition, one subset is meant for equally distributed classification accuracies among

all classes. The reason for this partitioning of available features lies in the structure of the hierarchical specialist selection strategy to applied as fusion technique and will be further explained in the following section. In order to discard irrelevant features, reduce the search space and to provide different feature sets (resulting in diverse ensemble members), we carry out feature selections. The n sets meant to support each of the n classes are generated by respectively labelling the supported class against the other categories in the data set. Resulting feature sets should be able to recognize associated classes with preference. The feature set aimed at balanced classification results undergoes the standard feature selection process without manipulation of labels.

The final recognition system is composed of a fusion strategy, which sits on top of the classifier ensemble. For this experiment we implemented four different fusion strategies. The classifiers in the ensemble are not restricted to a certain classification method, but can be chosen independently from each other. Due to the huge number of training samples we decided to start with the Naive Bayes classifier (Chapter 3.2.1), which is a simple classification scheme and extremely fast in training and test, even for high-dimensional feature vectors and large training databases. As feature selection method for ensemble generation we choose the computationally efficient correlation-based CFS method (Chapter 3.1.2).

The corpus provided by the INTERSPEECH 2010 Paralinguistic Challenge is divided into a training and development set containing 32 527 samples and 20 549 samples, respectively. For each sample the age in years and the gender (male, female or child) is given. Based on the age, samples are divided in seven different age groups (age groups 1-7), which are further aggregated into four general groups (child, youth, adult and senior). To maintain comparability with other submissions following a similar approach, we decided to use the set of standard features provided by the organizer composed of 450 prosodic features.

5.1.3 The Cascading Specialists Approach

Inspired by the ensemble-based systems described in Section 5.1.2 we decided to explore ensemble strategies and design a novel fusion scheme for classifier combination. We implement a hierarchical fusion approach that in contrast to (Lee *et al.* [108]) does not choose the order in a way that the easier classification tasks are found in the top levels of the hierarchy. Instead we first try to recognize difficult classes in order to gain overall accuracy by a well balanced recognition rate across classes. The algorithm is called the

Cascading Specialists (CS) approach and we implement several variations of the strategy. Performance is compared to standard fusion schemes and single expert classification.

Mean Rule (MEAN)

The mean rule is a standard decision level fusion method with an algebraic combination rule (Section 4.1), meant to combine the continuous outputs of all available ensemble members. By averaging the support given to each class ω_n the total support μ_n for class n throughout the whole ensemble can be calculated as:

$$\mu_n(x) = \frac{1}{T} \sum_{t=1}^{T} s_{t,n}(x) \tag{5.1}$$

T denotes the total amount of classifiers (therefore $\frac{1}{T}$ serves as normalization factor), $s_{t,n}$ describes the support given to class n by the t_{th} classifier in the ensemble. Finally the ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.

The appliance of this method is meant as a first appraisal of fusion potential and delivers some clues about compositions of the datasets, sample distributions and resulting characteristics in classification accuracies for single classes.

Cascading Specialists (CS)

In many cases, accurate investigation of confusion matrices for ensemble classifiers shows two phenomena concerning performance on single classes: A good true positive classification rate - which commonly serves as measure for single class performance - often goes hand in hand with a high false positive classification rate, because many samples are simply put into the inspected dominant class. In return classes with mediocre performance can sometimes also show low false positive rates, as a small amount of samples ever gets put into that category. The cascading specialists method bases on choosing specialists for single classes and brings them in a logical sequence in order to soften the mentioned phenomena. In a preparation step specialists for each of the n classes are selected by finding the classifier with best true positive rating for every class of the classification problem. These choices are based on evaluation of the training phase. Then the classes are rank ordered, beginning with the worst classified class across all classifiers and ending with the best one.

Given the preparation step and a test sample, the algorithm (Figure 5.2) works as follows: The first class in the sequence is chosen and the corresponding specialist is asked to classify the sample. If the output matches the currently observed class, this classification is chosen as ensemble decision. If not, the sample is passed on to the next weaker class and corresponding specialist whilst repeating the strategy. Sometimes the case occurs that none of the specialists classifies its connected class and the sample remains unclassified at the end of the sequence. Then the classifier with the best overall performance on the training data is selected as final instance and is asked to label the sample as ensemble decision.

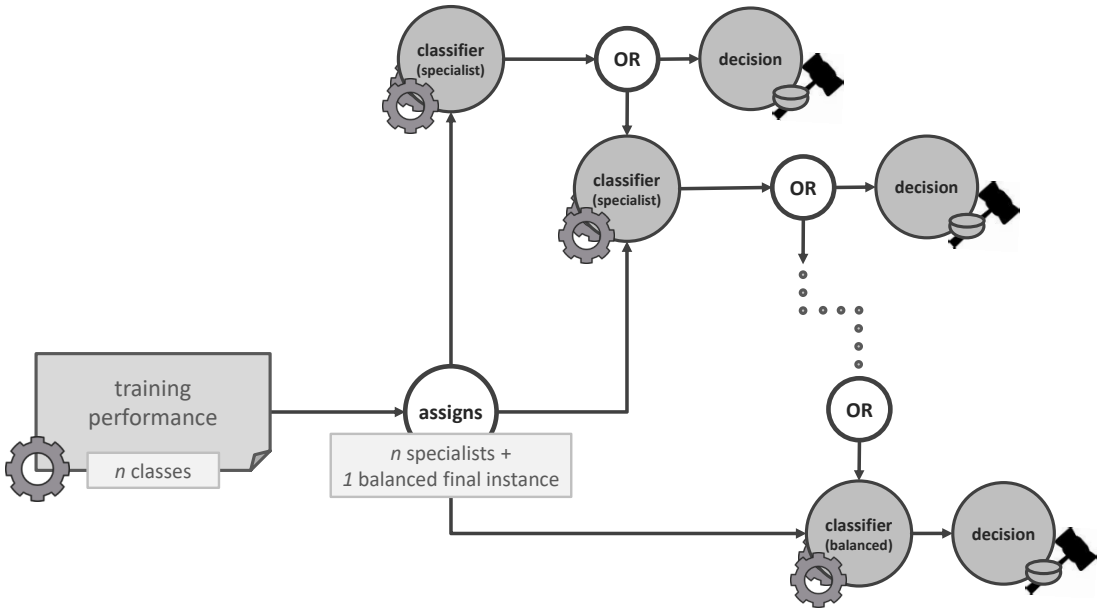


Figure 5.2: The hierarchical setup of the Cascading Specialists (CS) approach. Difficult classes are treated with priority in order to gain overall accuracy by a well balanced recognition performance across classes.

This strategy aims at a more uniformly distributed accuracy among classes. Weakly recognized classes are treated with priority and the belonging samples are more unlikely to end up falsely classified as a more dominating class later on. This results in a flattening effect that will at best improve overall classification performance.

The build-up of this ensemble fusion method explains the choice of sub-sets generated by feature selection described in Section 5.1.2. Feature sets supporting single classes are expected to serve as base for respective specialists while the single set covering all classes is meant to result in a classifier to be used for labelling samples which passed the process unclassified.

Cascading Specialists - One versus Rest (CS-OvR)

In the native CS approach, specialists are automatically chosen among a set of classifiers trained to recognize all observed classes. These specialists are rated based on their performance on the single classes. A more adapted approach can be developed by training classifiers specialised in separating their connected classes from $n-1$ remaining classes. They are no longer chosen based on evaluation of training performance but determined right from the start. This results in n classifiers dealing with binary classification problems. The established concepts of bringing the specialists in ascending order and intercepting unclassified samples with a well-balanced classification model are maintained.

The theoretical advantage of this strategy lies in the generation of best possible specialists for recognizing the associated classes. While this specialisation implies the potential of enhancing classification accuracies for single classes, there are some major drawbacks involved. The strict association of specialists with their belonging classes can lead to problems whenever training and test samples vary greatly, because there is no flexibility in specialist selection. This disadvantage also applies to the simple Cascading Specialists approach to a lesser extent, as its classifiers feature a broader classification potential than the one-versus-rest classification models.

Cascading Specialists - Multiple Specialists (CS-MS)

The last presented variant of the CS approach deals with the question of how to become less dependent on the evaluation of training data. A possible answer is to choose more than one specialist for every class and to execute internal fusion steps. This way the specialist assortment becomes more flexible and less conditioned by unreliable training performance, as the risk of choosing one wrong specialist becomes less important.

In detail $\lceil \frac{n}{2} \rceil$ classifiers among the ensemble are chosen to become specialists for each of the n classes. Whenever the sequence demands a new decision concerning class-belonging the corresponding specialists generate a decision via the weighted average rule. Again, the concepts of ascending class order and an intercepting classification model stay untouched.

The weighted average rule is an extension of the mean rule by adding weights to the classifiers and their probabilistic outputs.

$$\mu_n(x) = \frac{1}{T} \sum_{t=1}^{T} w_t s_{t,n}(x) \tag{5.2}$$

The formula is simply adjusted by the weight (deduced from overall performance on training data) w_t of the t_{th} classifier.

5.1.4 Comparison of Single Expert Systems to Standard and Custom Fusion Approaches

Table 5.1 shows the results obtained for experiments carried out with the Naive Bayes classifier. It describes different label-assignments, i.e. GENDER classifies the given samples in three classes describing probands as children, male or female adults. AGE4 groups them into four subsequent age-groups and AGE7 further sub-divides these partitions. Respective single classifier results are presented without (NO) and with (SEL) suitable feature selection and can be interpreted as baseline results. The following lines sum up the applied fusion experiments and show possible gains in classification accuracy compared to these baselines. Classification performance is given as unweighted average accuracy (UA) as described in Chapter 3.2.5.

Throughout all experiments decision level fusion methods (especially CS with multiple specialists - CS-MS) tend to outperform the single expert systems by about 5% without feature selection and around 1% with feature selection, whilst establishing a more balanced accuracy distribution among the single classes. The native cascading specialists approach (CS) performs stable on all tested datasets, as the aim of supporting under-represented or weakly recognized classes is nearly always achieved. A gain in classification accuracy for these classes of course is attained at the expense of stronger classes. Unfortunately the One-versus-Rest (CS-OvR) variant pushes this behaviour beyond justifiable limits, so that other classes accuracies get lowered in a disproportionate way. While the CS scheme mostly lead to enhanced overall performance - because of the appreciated flattening effect - the CS-OvR approach tends to lower the decision level fusion result in comparison to the baseline. Both methods suffer from disparities between training data and actual test data as only one specialist is chosen for respective classes and the selection of classification models as well as their structure heavily depends on evaluation of training. For example the similarities between child and female classes in the GENDER experiment seem to provoke the selection of wrong specialists. Furthermore the child class is heavily under-represented in the dataset. This fact together with implicit inhomogeneous characteristics of child voices (e.g. as the class sums up individuals before and after puberty vocal change) bears special risks for wrong specialist-selections between female subjects and children.

	NO	SEL	MEAN	CS	CS-OvR	CS-MS
GENDER						
male	90.20%	89.50%	89.20%	86.50%	87.80%	88.00%
female	61.00%	70.50%	75.30%	79.00%	62.10%	74.00%
child	47.30%	51.90%	49.50%	49.80%	64.30%	52.10%
UA	66.20%	70.60%	71.30%	**71.80%**	71.40%	71.40%
AGE4						
child	58.60%	63.40%	62.10%	50.10%	47.70%	62.30%
youth	13.00%	20.60%	26.60%	48.00%	57.70%	23.90%
adult	67.80%	44.90%	43.00%	25.20%	10.40%	40.20%
senior	8.20%	36.20%	37.20%	40.80%	45.40%	43.50%
UA	36.90%	41.30%	42.20%	41.00%	40.30%	**42.50%**
AGE7						
age-group-1	32.60%	41.90%	41.80%	46.50%	31.10%	44.20%
age-group-2	63.00%	63.20%	62.40%	56.30%	1.30%	56.40%
age-group-3	27.80%	45.40%	42.80%	50.60%	2.20%	45.90%
age-group-4	20.10%	26.40%	35.80%	30.60%	30.40%	34.80%
age-group-5	58.00%	32.20%	36.00%	17.50%	6.10%	30.40%
age-group-6	5.80%	19.10%	11.80%	23.20%	57.40%	19.00%
age-group-7	18.70%	29.30%	27.90%	33.70%	80.80%	29.60%
UA	32.30%	36.80%	36.90%	36.90%	29.90%	**37.20%**

Table 5.1: Evaluation of the AGENDER corpus sub-challenges GENDER, AGE4 and AGE7. Throughout all experiments hierarchical decision level fusion methods outperform the single expert systems.

The strategy of choosing multiple specialists for each class lowers the mentioned risks. It results in the most beneficial flattening effect of the three presented variants and therefore leads to best overall performance in most cases. Being less reliant on evaluation of training data further affirms the impression of multiple specialists being a good choice across given datasets. Overall, the intended uprating of under-represented classes can be achieved and the resulting flattening effect increases overall classification performance compared to single expert results obtained by the same classification model. All in all the cascading specialists fusion methods yield reliable classification rates with the huge benefit of a very balanced accuracy on all observed classes. They are capable of very precise detection for weakly recognised emotion-classes and therefore they can always be considered as possible fusion method for upcoming experiments.

5.2 Standard and Emotion-adapted Fusion Approaches for Bi-modal Data

The common recognition strategy for a classification problem with discrete labels obviously is to train all classifiers and corresponding ensembles to decide and label among one of the available classes. But if we for example take a look at the two-dimensional valence-arousal emotion model (Chapter 2.1), more emotion-adapted techniques for finding an ensemble decision can be proposed. The underlying idea is as simple as promising: Two or more ensembles are built and trained to recognise the observed emotion's axial alignment. Resulting outputs are combined later on for final decision. To compare the capabilities of emotion-adapted fusion approaches presented in (Kim and Lingenfelser [97]), we implement several standard fusion systems - ranging from very common voting mechanisms to rather seldom found fusion strategies featuring ranking algorithms and lookup tables.

5.2.1 Bi-modal Affective Data

For experimenting with these ideas we use a suitable dataset (Kim *et al.* [98]) that features four classes that base on the axes of the valence-arousal model. It is generated in a Wizard of Oz scenario by a modified version of the popular TV-show "Who wants to be a millionaire?". The participants take part in a quiz, interacting with a virtual quizmaster that is manipulated from outside to elicit the desired affective states throughout four phases of the experiment. Each resulting emotional class represents a quadrant (i.e. one of two possible extrema on the valence or arousal axis) and the respective prototypical emotion associated with it:

- **Pleasure** (low arousal and positive valence induced by easy questions and a fair quizmaster)
- **Joy** (high arousal and positive valence induced by difficult questions and a quizmaster that pretends most incorrect answers to be correct)
- **Boredom** (low arousal and negative valence induced by moderate difficulty and an annoying quizmaster)
- **Anger** (high arousal and negative valence induced by very difficult questions and a quizmaster misunderstanding correct answers)

Observed bi-modal data consists on the one hand of recorded speech data (SPE) of participants. Calculated features statistics over Mel-frequency cepstral coefficients (Imai [89], Krishna Kishore and Krishna Satish [102]). On the other hand, combined physiological measurements (BIO) of users are recorded within various channels: Blood volume pulse (BVP), respiration (RSP), skin conductivity (SC), electromyogram (EMG), and body temperature (TEMP). Statistical features are gained by analysing the the channels in time and frequency domain. Recognition results (Table 5.2) are gained on these modalities with SBS feature selection and a subject independent leave-one-user out cross-validation (Chapter 3). They serve as a baseline for the following ensemble experiments. Approaches not producing competitive results to the best available modality (SPE) may not be considered for practical application, as the use of a single expert system may be beneficial in that case.

	Single Modality Accuracy						
	BVP	**RSP**	**SC**	**EMG**	**TEMP**	**BIO**	**SPE**
Joy	29.51%	13.11%	0.00%	3.28%	13.11%	44.26%	32.79%
Anger	34.18%	29.11%	27.85%	0.00%	0.00%	43.04%	58.23%
Boredom	44.76%	54.29%	51.43%	71.43%	75.24%	51.43%	71.43%
Pleasure	52.04%	35.71%	47.96%	62.24%	28.57%	59.18%	54.08%
UA	40.12%	33.06%	31.81%	34.24%	29.23%	49.48%	**54.13%**

Table 5.2: Recognition results for single modality classification. The speech modality (SPE) achieves best results and serves as baseline for the following ensemble experiments.

5.2.2 Standard Fusion - Voting, Ranking and Lookup Tables

Fusion strategies in Chapter 4.1.3 describe the most widely used algorithms to combine the outputs of several ensemble classifiers on decision level. In addition to the emotion-adapted approaches, we implement ranking and lookup table approaches to class label combination that are not commonly found in ensemble studies and compare them to basic voting mechanisms as well as emotion-adapted fusion strategies.

Voting

The perhaps most obvious and basic approach to the intended combination of classification models is to provide each ensemble classifier (e.g. one classifier per modality

within the multi-modal system) with a vote (Figure 5.3). This vote is associated with the class-label that was predicted by the classifier. Then a classical polling can be carried out. Several ways of evaluating the cast votes come into consideration.

The easiest way to hold a poll is to give every ensemble member a vote a value of one. Afterwards the votes are enumerated and one of the following methods could be adopted to bring about a final ensemble decision. Unanimous voting demands all classifiers to vote for the same class. If not defined otherwise, no ensemble decision is generated every time one or more ensemble members disagree. This characteristic makes unanimous voting near useless for practical usage, because it is very unlikely that the whole ensemble ever completely agrees on one single class-prediction. In simple majority voting a final ensemble decision is attained if one prediction is able to sum up an absolute majority of votes. This means that at least one more than the exact half of the ensemble's spendable votes are dedicated to the same prediction. Though this criterion is not as severe as in unanimous voting, it still is too restricting for usage in the field. In majority voting all classifiers freely distribute their votes among the eligible class-labels. The ensemble decision is simply determined by finding the class-label with the most obtained votes. This approach to decision making is non-restraining enough to practically apply it to actual ensembles.

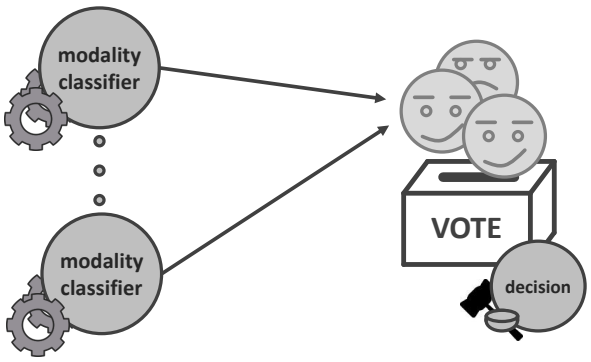


Figure 5.3: Simple voting schematic.

Contrary to these simple voting mechanisms a weighted voting approach doesn't put the same emphasis on any given vote. It considers the fact that not all classifiers within the ensemble perform equal or at least comparable. Some trained classifiers may unfortunately be underperformers, some may shape out to be real specialists for a certain classification problem. These imbalances in classification performance can be taken into account for finding the ensemble decision. Subsequently the question arises where the prior knowledge should stem from. Even if previous experiments on the relevant ensemble are at hand for evaluation and performance forecasting, this knowledge should

be used with caution. A momentarily classified sample may not be associated with prior knowledge that was gathered using this sample. Otherwise the classification results become questionable as too much information about the classified samples are available to the classification model. As in the Cascading Specialists approaches presented in Section 5.1.3, the weighting of ensemble votes relies on evaluations on the training data. Two forms of weighted majority voting have been analysed.

The average weighted majority voting method does not give every voter's choice an equal emphasis. Instead the output of each classifier is weighted by the overall performance on the training data and therefore well performing classifiers are taken into account for the poll with greater influence.

For class weighted majority voting, we assume there can be experts for certain classes in the ensemble. The classifier votes are consequently weighted by the classifiers training data performance on the class it chooses. This proceeding backs up classifiers with low average performance but good qualities in detecting a certain class.

	Standard Fusion I.	
	Average Weighted	**Class Weighted**
Joy	19.67%	3.28%
Anger	32.91%	12.66%
Boredom	78.10%	88.57%
Pleasure	62.24%	57.14%
UA	**48.23%**	40.41%

Table 5.3: Recognition results for majority voting with averaged and class based weighting methods. The latter approach suffers from overemphasising of strong classes.

Both weighting methods resemble an automated selection of the better channels for decision making. Average weighting outperforms the approach of class weighted voting with significant values, which mainly results from the observed overemphasising of strong classes in the latter approach. All in all the mediocre performance of the single physiological channels drag down the the performance of the ensemble, as simple voting logics cannot shape out the advantages of strong ensemble members like the speech modality (SPE) or the combined physiological channel (BIO). This finally results in worse performance than a single expert system (Table 5.2). Better solutions to fusing ensemble decisions therefore need to be explored within the upcoming experiments.

Enhanced Borda Count

The Borda count (BC) algorithm can best be described as a ranking mechanism. According to (Polikar [135]), it was first developed by Jean Charles de Borda and is a commonly used practical voting application. A good example would be the Eurovision Song Contest, where every country rank orders the contesting songs and points are distributed accordingly. The genuine form of BC is a special form of simple voting and it theoretically features several advantages. During the ranking process every classifier chooses the most probable of n classes and assigns it $n-1$ votes. Then the remaining classes are ranked downward and given one vote less every step until the least ranked class is provided with zero votes. Finally the overall votes for every class are added up and the class with the most support is chosen as the ensemble decision (Figure 5.4).

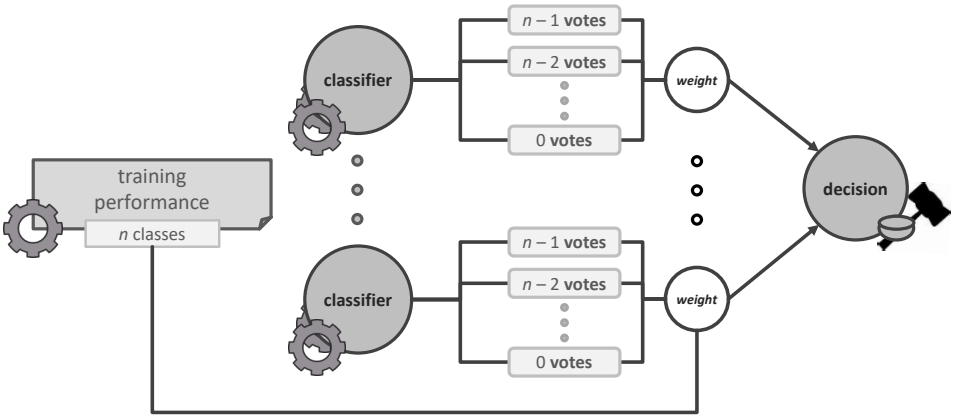


Figure 5.4: The enhanced Borda count ranking mechanism as fusion method.

The described ranking algorithm demands a special requirement from the used classification models. An applied classifier must not only be capable of producing an output shaped as single class label, but has to have the potential of ranking all class-candidates. The normalized probabilities for every single class can be rank ordered and then serve as basis for vote distribution.

By following this procedure, voting ties are very unlikely to happen and the choices of not winning voters are not thrown away but are included in the decision. The second mentioned advantage is of special interest, cause the feature of considering the support a classifier gave to a not winning class is very uncommon in decision level fusion and could have a positive effect on the ensemble's overall classification accuracy. Of course one can try to improve the results of BC by adopting weighting techniques from former voting mechanisms by altering the weight of the provided votes with either average weights or class weights and therefore enrich the ranking mechanism with the according

advantages of weighted majority voting. This results in two possible forms of enhanced BC, depending on which weighting method is chosen.

	Standard Fusion II.		
	BC	**Enhanced BC (1)**	**Enhanced BC (2)**
Joy	8.20%	13.11%	0.00%
Anger	25.32%	27.85%	0.00%
Boredom	76.19%	76.19%	93.33%
Pleasure	60.20%	65.31%	40.82%
UA	42.48%	**45.62%**	33.54%

Table 5.4: Recognition results for the ranking algorithm Borda count with and without enhancement through averaged (1) and class based (2) weighting.

This special approach to voting containing a ranking mechanism delivers in standard form a barely acceptable set of results. These are consistently upgraded by the appliance of the average weighting method, but they always stay behind former tested voting methods. The drawbacks of class weighting (2) appear in the most evident form. The disregard of weak classes increases in effect so that classification accuracies concerning these classes very often drop to zero. The usage of the class weighting method for the described voting mechanisms can now be definitely dis-advised.

Behaviour Knowledge Space

Another way to reasonable decision making within an established ensemble is the construction of lookup tables. This topic and corresponding algorithms are proposed and discussed (Huang and Suen [88]) under the term behaviour knowledge space (BKS). The underlying idea focuses on the various combinations of labelled predictions that the ensemble's classifiers produce during the training phase. Given a sample to be classified for training, each member of the ensemble computes a class prediction and provides it via the predicted class label. Depending on the different forecasts of the classifiers, varying combinations of labels can be observed. The monitored combinations are provided with the belonging true class of the sample (which of course is known during training) and stored in a lookup table. After the training has ended, equal combinations are summed up and finally get associated with the true class, which the considered label combination is meant to represent. If two or more identical combinations of class-predictions were generated for different true-classes, the true class that was attached the most often to this

label combination gets chosen. Table 5.5 illustrates an excerpt from a readily trained lookup table for a four-class classification problem with classes c_i with $(i = 1; 2; 3; 4)$ that is investigated by an ensemble formed by two members e_1 and e_2.

Trained BKS Table		
Prediction e_1	**Prediction** e_2	**Actual Class** c_i
1	1	c_1
1	2	c_1
1	3	c_3
1	4	c_3
2	1	c_2
2	2	c_2
...	...	...

Table 5.5: Exemplary behaviour knowledge space lookup table for a four class prediction problem with two ensemble members.

One may recognise that some true class c_i possibly gets associated with a label combination that does not contain the class. This phenomenon is legitimate and welcome as it obviously describes one possible, characteristic behaviour of the ensemble when facing a sample of this class. Classification can now be executed by requesting the predictions of the ensemble and comparing the resulting label vector to the lookup table. The actual class, which the table associates with the given combination, is chosen as ensemble decision.

This technique has one major disadvantage. If training data is not dense enough not all possible labelling combinations are covered by the knowledge space and several test sample combinations may not be classified by it. This problem can be avoided by the construction of further smaller lookup tables. The classifiers with the worst remaining overall performances on the training data set are successively left out. If a test sample combination cannot be classified with the current lookup table it is compared to the next smaller whilst ignoring the according entry in the test sample combination until a fitting lookup table is found. The enhanced algorithm can be broadened and optimized by a step in the training phase that estimates the most promising size for the biggest table to be constructed.

The surprisingly bad results of standard BKS is the first fact that catches the eye. This shortcoming simply stems from the amount of labelling combinations that are not found

	Standard Fusion III.		
	Standard BKS	**Enhanced BKS**	**Optimized BKS**
Joy	6.56%	34.43%	31.15%
Anger	16.46%	49.37%	60.76%
Boredom	32.38%	66.67%	61.90%
Pleasure	31.63%	57.14%	55.80%
UA	21.76%	51.90%	**55.80%**

Table 5.6: Recognition results for standard, enhanced and optimized versions of the BKS lookup approach.

by the original lookup table during the training phase. Belonging samples just pass through the algorithm unclassified and result in effectively useless classification accuracy. The further establishing of the enhanced BKS completely prevents unclassified samples and generates results on a competitive level in comparison to voting approaches while working with the limited number of training samples available. One can positively notice the balanced classification accuracies among classes, overemphasising of dominant classes does not seem to be a problem. The best performance is achieved by optimized BKS. Typically starting with a lookup table covering the three winning classifiers of the ensemble, it generates slightly superior classification accuracy than enhanced BKS, which always starts with a lookup table of size seven.

5.2.3 Emotion-adapted Fusion Approaches

We have by now seen the performance of simple voting mechanisms as well as more sophisticated implementations of ranking mechanisms and lookup tables for bi-modal affect recognition. The question however is, if the so far achieved recognition performance can be elevated if we tune the fusion algorithm to the emotion model given by the affect classification problem. An approach researching into the fragmentation of emotions into valence and arousal has been investigated by Kim and André [95] and lead to convincing amendments in classification accuracy on the treated dataset. Kim and André [95] introduce the nomenclature *dichotomous* for recognition approaches that handle emotion model axes instead of discrete labels. In this case, the term expresses the dyadic decomposition of the classification problem using valence and arousal. Because of the mentioned mapping to orientation in the two-dimensional emotion model these algorithms can no longer be generalised for any classification problem and we refer to them as emotion-adapted. The applied two-dimensional emotion model is well suited for

binary disassembling of samples into emotions with high or low arousal, respectively positive or negative valence. In contrast to direct classification, immediate mapping onto one of the four desired emotion classes is not possible but an additional combination step is needed for final decision.

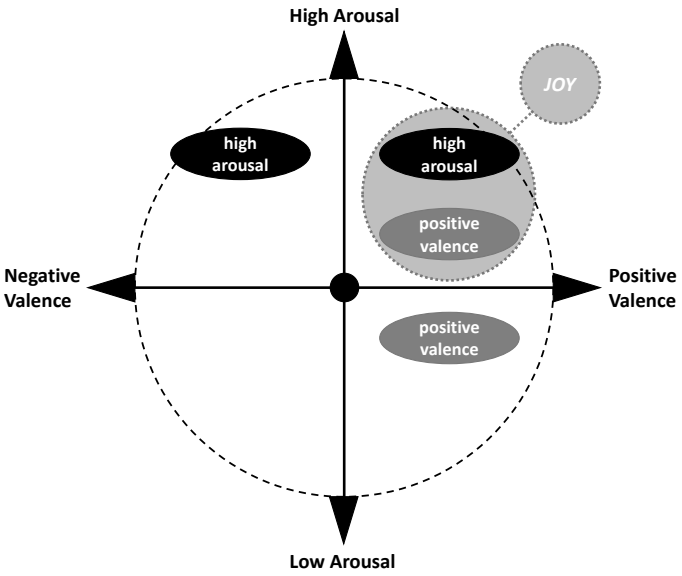


Figure 5.5: Exemplary decomposition of emotion joy into arousal and valence orientations.

As a start two autonomous ensembles are established to estimate arousal and valence orientations. All available channels are considered and the samples available for training must be re-labelled. The observations contained in the dataset are no longer associated with a concrete class membership but get tagged with the corresponding grade of arousal and value of valence. Because of its good performance in the first experiment, the cascading specialists approach (Section 5.1.3) is adopted for decision making within the gained classifier ensembles. This fusion method has proven to be capable of producing good classification accuracies and well balanced results among single classes.

Static Combination and Lookup Tables

After two accurate ensembles are trained and decisions can be made via cascading specialists, one ensemble serves as expert for classification of positive or negative valence, the other one is used to decide whether the examined emotion features a high or a low grade of arousal. Having obtained labelling outputs from both ensembles the combination of positive or negative valence with high respectively low arousal allows a clear assignment to one of the four emotional states featured in the 2D emotion model.

This predefined, static class-designation is genuinely deduced from the emotion model and strictly follows the rules illustrated in Table 5.7.

Static Combination Rule		
	high arousal	**low arousal**
positive valence	*Pleasure*	*Joy*
negative valence	*Anger*	*Boredom*

Table 5.7: Static combination rules for combining outputs of valence and arousal classifiers.

The static combination step finally leads the classification task back to the original four class problem. The final label-outputs can be evaluated as usual and again be compared to former fusion methods.

The dichotomous approach is well suited for further expansion. Two possible enhancements initially come into consideration: The method of static combination is not flexible and the combination rules are not naturally evolved from the classification problem during training but are dictated and supervised from outside. A more adjusting combination step would be preferable and in addition the usage of supplementary ensembles should become possible. In former approaches lookup tables were used to keep track of the outputs of single ensemble classifiers. This function can easily be altered to chart the decisions of several ensembles and evaluate them in previously described manners. Through this new form of combination supervised combination rules become obsolete and the first related experiment investigates the supplementation of arousal and valence ensembles with a lookup table utilising their decisions.

As a lookup table organised to handle ensemble decisions is able to track more than two ensembles, further information can be brought into account via an accessory ensemble. Direct classification via the cascading specialists method generates an additional entry in the lookup table proposing one of four possible classes. This enhanced approach leads to a knowledge space covering decisions of three ensembles. Available training data turns out to be dense enough to use only one single table, otherwise the developed methods for enhanced BKS (Section 5.2.2) could be applied.

Table 5.8 shows recognition accuracies for emotion-adapted ensembles using the static combination rule, lookup tables for valence-arousal decisions (1) and lookup tables combining valence, arousal and direct decisions (2). The results of the lookup table monitoring outputs of arousal and valence ensembles exactly match the results of static combination. This observation confirms the correctness of the predefined combination

	Emotion-adapted Fusion I.		
	Static Combination	**Lookup (1)**	**Lookup (2)**
Joy	34.43%	34.43%	32.79%
Anger	54.43%	54.43%	51.90%
Boredom	70.48%	70.48%	76.19%
Pleasure	57.14%	57.14%	64.29%
UA	54.12%	54.12%	**56.29%**

Table 5.8: Recognition results for emotion-adapted fusion using the static combination rule, lookup tables for valence-arousal decisions (1) and lookup tables combining valence, arousal and direct decisions (2). Results of the lookup table monitoring outputs of arousal and valence ensembles exactly match the results of static combination.

rules deduced from the emotion model. On the other hand the additional effort of establishing a knowledge space obviously is not necessary when dealing only with information drawn from arousal and valence. The knowledge space containing additional information concerning direct classification provides advanced classification accuracies for one single subject and the subject independent dataset (the overall performance on this dataset is the most accurate one recorded so far). Additions of references to concrete classes are apparently capable of enhancing the binary decisions based on valence and arousal but do not necessarily influence them in a beneficial way. The knowledge space treats every ensemble decision with equal emphasis. If prognoses on direct class membership are meant only to support the dichotomous decision, another form of combination has to be developed.

Dichotomous Voting Approaches

The dichotomous voting approaches to be introduced heavily rely on emotion theory. Combination of proposed values for valence and arousal serve as basis for decision making. Afterwards the prognosis given by the ensemble responsible for classification of direct class membership is considered for an alteration of the present decision. The vote of direct classification is able to reinforce the arousal-valence combination or to shift the chosen emotion to an adjacent quadrant of the two-dimensional emotion model. Figure 5.6 illustrates an example decision stemming from valence-arousal classification and its possible alterations through tendencies to concrete classes.

Note that this model does not give direct classification influence the possibility to alter the valence-arousal decision to a complementary emotion. This behaviour is intended

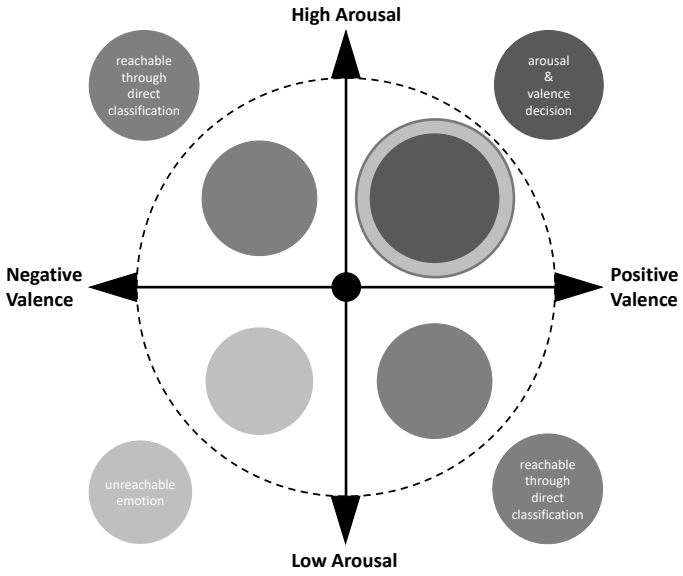


Figure 5.6: Reachable quadrants of direct tendencies.

and meant for stronger emphasising on dichotomous classification than on direct class allocation. The appropriate algorithm implementing this abstract model utilises proper vote distribution among the possible emotion-quadrants. Valence and arousal ensembles give their votes of numerical value 1 to both quadrants that accord with their orientation, resulting in a base distribution of $(2-1-1-0)$ among the four quadrants. By providing the choice of direct classification ensemble with a value of 1.5, the base decision can be raised to 3.5 or one of the adjacent quadrants is able to outnumber the base decision by rising to a value of 2.5. The quadrant that was given zero votes by dichotomous decision can never be chosen as final decision, because even if direct classification should chose it, a $(2-1-1-1.5)$ distribution still favours the base decision.

Performance of direct classification still seems to be very crucial for the whole approach as it finally decides between three possible classes. But obviously the overall accuracy may rather worsen because direct classification generally performs not as well as dichotomous approaches, so another form of information backing up dichotomous decisions could turn out to be more effective. The concept of a cross axis is meant to provide additional information to dichotomous approaches. This axis can not be directly deduced from emotion theory but from a mathematical point of view it is a reasonable partition of the two-dimensional model. Just like valence and arousal axes the cross axis divides the emotion model into two separate parts, each containing two emotion-quadrants. These parts contain the respective, complementary quadrants and therefore split the model in a diagonal way. The resulting halves contain the contrary emotion-classes Joy and

Boredom as well as Anger and Pleasure respectively. Table 5.9 shows in recognition numbers that the cross axis partition leads to reasonable classification results compared to the valence and arousal partitions legitimated by emotion theory.

	Dichotomous Ensembles		
	Arousal	**Valence**	**CrossAxis**
UA	74.89%	73.27%	71.14%

Table 5.9: Recognition results for valence, arousal and cross axis partitions. Though not legitimated by emotion theory, the cross axis model leads to reasonable classification results.

We can now take advantage of all introduced ensembles in a more refined voting mechanism: Arousal, valence, cross axis and direct classification. Every listed ensemble generates its decision via cascading specialists method. The provided votes are all given a numerical value of one and then take part in a stepwise combination process positively leading to a final decision. Classification is guaranteed, as the final step inevitably leads to a result (if preceding steps could not establish it due to voting ties).

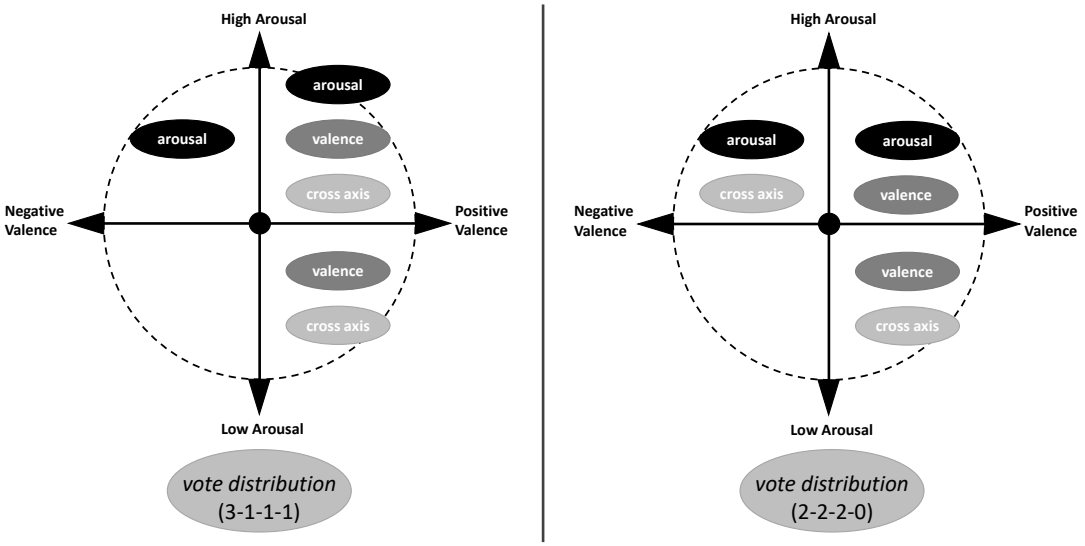


Figure 5.7: Step1: Possible vote distributions.

Step 1: Combination of Arousal, Valence and Cross Axis This step exactly matches the static combination method presented in dichotomous approach with cross axis. Each ensemble distributes its votes among the two quadrants that fit the recognised alignments in the 2D emotion model. This step results in one of two possible outcomes (Figure 5.7):

(3-1-1-1) If the ensembles agree on one emotion-quadrant, it recieves three votes and can already be chosen as final decision.

(2-2-2-0) If the ensembles do not manage to agree on one emotion-quadrant, a voting tie occurs. No final decision can be chosen, instead the draw has to be dissolved and the algorithm moves on to the next step.

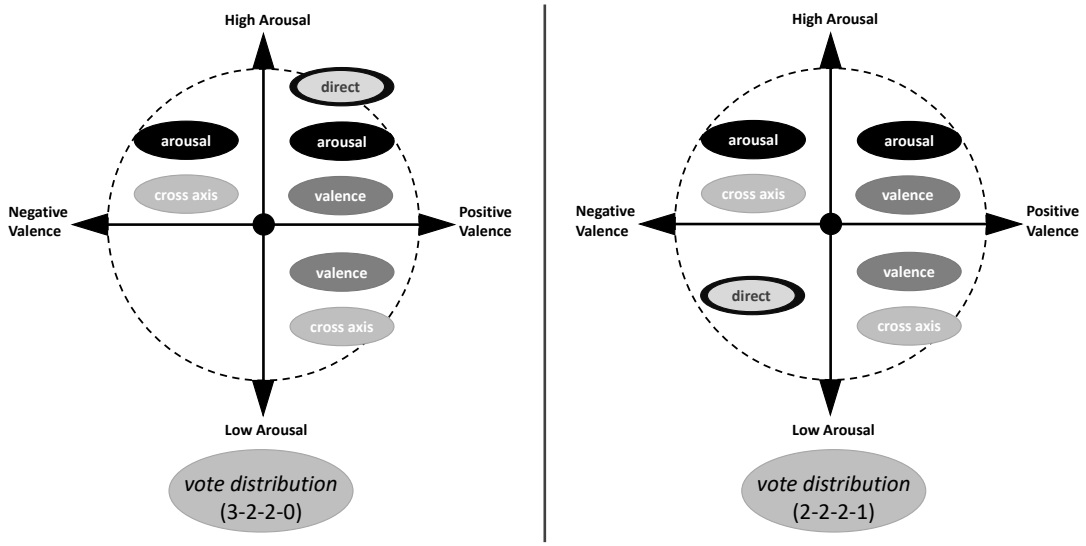


Figure 5.8: Step2: Possible vote distributions.

Step 2: Resolving of Draws through Direct Tendencies In order to resolve the draw, the direct classification ensemble designates exactly one vote to the class it predicts. Two situations can arise through this supplemental vote (Figure 5.8):

(3-2-2-0) If the ensemble chooses an emotion-quadrant that already holds two votes, the tie is resolved and the corresponding emotion is determined to be the final decision.

(2-2-2-1) If the ensemble chooses the emotion-quadrant that has not recieved any votes yet, the tie is not resolved and the last possible step has to be executed.

Step 3: Decision through Arousal and Valence Combination If actually no decision could be established by the previous steps, the emotion-class that was originally determined by arousal and valence ensembles is ultimately chosen as final decision. In

practice this case rarely occurs, but it is definitely needed to guarantee that no sample passes the decision process unclassified.

	Emotion-adapted Fusion II.	
	Voting without CrossAxis	**Voting with CrossAxis**
Joy	42.62%	49.18%
Anger	56.96%	55.70%
Boredom	64.76%	70.48%
Pleasure	58.16%	66.33%
UA	55.63%	**60.42%**

Table 5.10: Recognition results for emotion-adapted fusion.

All in all results seem to improve with every additional ensemble making the emotion model more accessible, which leads to the definition of a cross axis ensemble. The success of dichotomous voting without cross axis could be described as mediocre. Classification performs on an acceptable level, careful observation of the algorithm's steps for single samples shows that in the latter cases correct base decisions by valence and arousal combination are too often overwritten by wrong direct classification votes. However, gained results of the voting steps including the additional cross axis ensemble lead to the best recorded results throughout this experiment. This achievement is satisfying and legitimates the augmented complexity originating from usage of four ensembles and the needed combination steps.

5.3 Conclusions about First Experiments

Observation of many possible fusion experiments has proven decision level fusion to be capable of interesting improvements in classification accuracies compared to single expert systems. Additional effort is always necessary compared to standard classification, but the gains definitely seem to be worth it. Most generalised methods do not lead to major improvements concerning classification rates, but all presented advantages of ensemble appliance are entirely provided: Multi sensor applications relying on these procedures can establish a certain robustness against breakdowns of one or more sensory devices. Training efficiency is guaranteed through partitioning into subsets of huge datasets or re-sampling of small datasets for different classifiers vice versa. Following the divide and conquer theory, complex feature distributions can be learned more effectively

by ensembles than by single classifiers. Perhaps the most important benefit is the very adequate possibility to combine multiple different modalities via appropriate channels, classifiers or ensembles. Furthermore generalised methods - especially voting methods - can be implemented in a straightforward way, as they do not need closer understanding of underlying datasets, assembling of classes or ensemble theory.

Well balanced results and good overall performance were achieved with the custom cascading specialists (CS) approach. Other proposed standard fusion approaches can be traced back to known methods, though they were consistently adjusted and enhanced to better suit the given scenario. At this point the possible potential of behaviour knowledge space (BKS) can be mentioned. This approach surely generates very acceptable results if huge datasets are available for training. Unfortunately the emotion dataset building the base for present experiments does not provide enough samples to fully investigate its usefulness.

Emotion-adapted approaches lead to far better improvements than standard fusion strategies, though deeper understanding of the applied emotion model is necessary. Also more than one ensemble and newly developed voting rules are needed in order to achieve the possible benefits. Generally said, the more ensembles are generated for broadening information coverage concerning the given dataset and class distribution, the better the gained improvements in classification performance. Appraisal of justifiable effort against possible advantages is a trade-off decision that has to be made individually for every new application.

Chapter 6

A Systematic Discussion of Synchronous Fusion on Natural and Acted Datasets

In the emerging stages of affect recognition as a scientific field, most effort was put into single-channel emotion recognition, while only a few studies with focus on the fusion of multiple channels have been published. Even though most of these studies apply rather simple fusion strategies - such as the sum or product rule - some of the reported results showed promising improvements compared to the single channels. We have seen in first experiments that a user's emotional state can naturally be perceived by combining emotional cues derived from available affective channels. Such results encourage investigations if there is further potential for enhancement if more sophisticated methods are incorporated. In Chapter 5 we have only applied fusion on the decision level and herein mostly chose class label combination strategies. To broaden our insights in the capabilities of fusion algorithms in multi-modal affect recognition, we need to greatly extend the number of implemented fusion strategies and compare their performance. In (Lingenfelser *et al.* [110]), we apply a wide variety of possible fusion techniques such as feature fusion, decision level combination rules, score level fusion and a hybrid approach. We carry out a systematic comparison of a total of 16 fusion methods on different corpora and compare results using a novel visualization technique. The goal of this extensive comparison is twofold: First, we search for indication if there is a generally advisable combination rule, that based on its combination logic or level of appliance has a clear edge on other fusion methods. Secondly, we want to compare the performance of fusion methods across different corpora. A good result on one corpus may not transfer to

another dataset and so no general advise can be given. Special attention is hereby given to differences between natural and acted datasets. While the latter ones may be useful for identifying good feature sets and establish a greater amount of training data, the final goal is to make affect recognition systems generalisable in natural applications.

6.1 No Free Lunch

Basing predictions on emotional states on multiple channels demands the fusion of multi-modal observations at some point of the prediction process. As described in Chapter 4 this effort can be done at different levels - most commonly on the feature level by merging cues from all modalities into one classification scheme, or at decision level by combining outputs of several classifiers. If we however consider the great amount of meanwhile established and further possible ensemble-based strategies, the question arises if there exist generally advisable ones or if the success of a strategy is based on the observed problem.

The *No Free Lunch* theorem (Wolpert and Macready [181]) has proven for supervised machine learning that there is no universally applicable classification scheme for all given classification tasks. When observing all possible problems, solutions perform on an equal level on average. Studies on multi-modal combination techniques such as Duin and Tax [51], Kuncheva [104, 105] or Fumera and Roli [68] examine synchronous fusion strategies and sometimes advise on which scheme dominates others. Results are not consistent throughout mentioned experiments, so suspicion that *No Free Lunch* holds for combination rules as well as for the underlying classification methods seems reasonable.

In the field of affect recognition fusion has by the time of the following study been mainly applied to audio-visual data. Zeng *et al.* [187] cite 18 studies dealing with audio-visual fusion. The authors distinguish between feature, decision and meta level fusion, where the latter is identical to score level fusion, which use a 2nd-level classifier to combine predictions of the single channels (Chapter 4.1.4). While none of the mentioned studies uses methods of all three kinds, it is also difficult to compare the results between the studies as they differ greatly in their methodology, as well as the underlying databases. We will now try to enrich the ongoing discussion with a comprehensive comparison of various established and novel fusion strategies, ranging from feature fusion and elaborated decision level combination rules to score and hybrid level fusion. These will be applied to different (natural and acted) corpora for emotion recognition in order to directly compare relative recognition success on different classification problems. We hope to give clear

hints on benefits of certain fusion schemes or even their interchangeability for future studies in the sense of *No Free Lunch*.

6.2 Affective Corpora: Natural and Acted Datasets

Within the review and comparison of multi-modal affect recognition studies, D'Mello and Kory [48] also include whether the published results were gained on natural or acted datasets. The multi-modal effect, which describes the relative gain of a multi-modal fusion approach compared to best single channel performance (Chapter 4.2.1), is about three times higher on acted datasets (4.4% improvement on natural data compared to 12.1% on acted). Though mentioned survey was published after the study presented within this chapter, performance differences of machine learning and ensemble approaches between natural and acted data were expected. Consequently two different corpora are used to draw a comparison of available fusion techniques - the DaFEx (Battocchi *et al.* [8]) and the CALLAS (Caridakis *et al.* [26]) corpus. These corpora have been chosen as they both contain audio-visual recordings of Italians (the CALLAS corpus actually features more modalities and cultures). They differ, however, in the number of expressed emotional states and their level of naturalness.

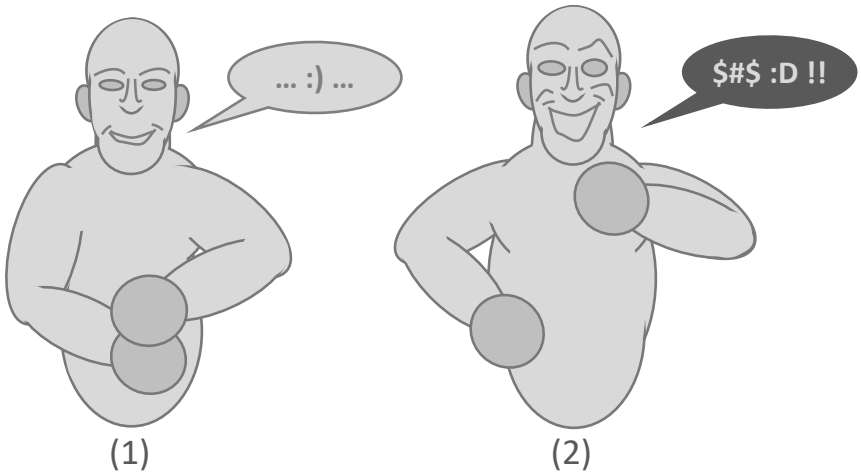


Figure 6.1: Acted emotions (2) tend to be prototypical and exaggerated, which could lead to a more reliable discriminability of resulting feature values. Natural emotions (1) are expressed individually and by more subtle cues, making classification and generalisation more challenging.

Studies dealing with the recognition of affective states are often based on recordings of acted emotions performed either by professional actors or amateurs. Figure 6.1 exemplarily depicts the sometimes very obvious differences between acted and natural

data: Acted emotions often are very prototypical and shown in an exaggeratted fashion, which could lead to a more reliable discriminability of features. Natural emotions on the other hand are expressed individually and by more subtle cues, making classification a challenging task. Classification systems trained on unnaturally expressive emotion samples may perform insufficiently in real world applications. Maybe because of these challenges we can by now observe an increasing trend towards more natural data sets. It turned out that findings derived from acted data can not necessarily be transferred to spontaneous emotions (Vogt and André [164]). Using corpora of both kinds allows us to investigate to what extent the choice of the fusion technique depends also on the naturalness of the observed data.

The DaFEx corpus contains recordings of eight professional Italian actors (four male and four female) expressing six basic emotions and neutral. It was initially constructed as a benchmark for the evaluation of facial expressivity of embodied conversational agents, but is well suited for the evaluation of emotion recognition systems, too (Rabie *et al.* [138]). It consists of 1008 short videos clips, where each clip corresponds to one the basic emotions happiness, surprise, fear, sadness, anger and disgust, or neutral. The facial expressions are available at three intensity levels (low, medium and high), but for our purpose are combined to a single class. Finally, we select only those samples, where the actors were also uttering a sentence, resulting in 84 samples per subject equally distributed among the seven classes.

The CALLAS expressivity corpus was constructed within the European Integrated Project CALLAS. Designed for the examination of cultural differences, it was recorded in three countries, Germany, Italy and Greece. In contrast to the DaFEx corpus, participants in the CALLAS corpus have no special acting abilities. An approach to elicit true emotional reactions is of course a lot more complicated to conduct than just having actors express the prototypical emotions.

During a single session 120 emotion inducing sentences were successively presented to the participants. The sentences are derived from the Velten mood induction technique (Velten [162]) and based on their semantic content, they can be divided into three categories: *Positive*, *negative* and *neutral*. After the silent reading of a sentence the projection was blanked out and the phrase was expressed by the probands in their own words and with whatever gesture or voice they felt to be fitting. Please note that recorded persons had no acting background and it was their personal discretion how and to what extent the emotion should be expressed. This setting leads to a more diverse set of displayed behaviour as it would otherwise occur under a more restrictive setup. However, it comes closer to what a system must expect under a realistic setting. What a recognition

system must deal with under real-life conditions is for example reflected whenever persons were not at all using gestures to accompany their speech - a fusion system should not *expect* information every channel.

Emotion	Sentence
positive	The hike was fantastic! You wonǵ believe it! But we made it to the top!
neutral	The names on the mailing list are alphabetically ordered.
negative	Sometimes I wonder whether my effort is all that worthwhile.

Table 6.1: Example sentences in three emotional categories.

While users were reproducing the mood inducing sentences, their behaviour was recorded with two cameras - one capturing at the the face and one the whole body. In the following comparison study we only analyse facial videos. Voice was captured from a fixed high quality microphone mounted above the users head. For this study we only consider the Italian sub-corpus which consists of 1539 samples of 13 persons (seven female and six male) equally distributed among the three classes. Specific emotions were elicited by a mood induction technique. Samples were consequently labelled corresponding to the state of the stimuli sentences.

6.3 Systematic Comparison

The comparison study we present in (Lingenfelser *et al.* [110]) carries out a comprehensive comparison of synchronous fusion approaches on the described corpora, using an equal modality treatment as in (Wagner *et al.* [168]): Like every typical machine learning pipeline, we start with the extraction of descriptive features to convert the raw signals into the compact form required for classification. We restricted the set of features to those that can be extracted in real-time and in a fully automatic manner. For example, no features have been included that require information on the spoken words or grammatical context, as such information is difficult to get without manual annotation. As in our first experiments, we extract acoustic features related to the paralinguistic message of speech: MFCCs and spectral features as well as prosodic features from pitch, energy, duration, voicing and voice quality total to the amount of 1316 features calculated by EmoVoice (Vogt *et al.* [166]). For video analysis we use SHORE, a library for facial emotion detection developed by Fraunhofer IIS (Küblbeck and Ernst [103], Ruf [142]). In the first place, SHORE offers a robust tracking of in-plane rotated faces up to sixty

degrees. For each face that is found, SHORE reports the bounding box of the face, as well as the position of the left/right eye and the nose tip. These features measure head movement. In addition, the left and right corner of the mouth and its degree of opening is reported. Most important, SHORE also calculates scores for four facial expressions – namely happy, angry, sad and surprised. These scores are also extracted for each frame and used in addition to the geometric features. In total, for each segment a series of 24 short-term features is derived by joining features extracted for each frame in the clip. Finally, we extract 11 long-term measurements, leading to an overall feature set with 264 entries (Table 6.2).

Modality	**Channels**	**Short-term Features**	**Long-term Features**	**Total**
Voice	Mono Audio	Pitch	Mean	1316
	48kHz	Energy	Median	
		MFCCs	Maximum	
		Spectral	Minimum	
		Voice quality	Variance	
			Median	
			Lower/Upper Quartile	
			Absolute/Quartile	
			Range	
Face	Video	Face Bounding Box	Mean	264
	25Hz	Eye Position	Energy	
		Mouth and Nose	Standard Deviation	
		Opening of Mouth	Minimum	
		Facial Expressions	Maximum	
		Happy	Range	
		Angry	Position	
		Sad	Number Crossings	
		Surprised	Peaks	
			Length	

Table 6.2: Overview of feature extraction methods applied to modalities audio and video in the comparison study. From the mono audio channel and video images we extract short-term features and compute statistical long-term features from these respectively.

Feature sets for both corpora are reduced by correlation-based feature selection followed by a sequential forward search (Chapter 3.1.2). After feature selection 35 audio and 40 video features remained for the DaFEx corpus, while on the CALLAS corpus 64 audio and 45 video features were chosen. All classification tasks within the comparison study are done via Naive Bayes (Chapter 3.2.1). We chose this rather simple classification scheme due to its successful application in earlier emotion recognition tasks (Vogt and

André [164]) and its fast computation, which allows us to run our experiments in a reasonable amount of time. For the evaluation of the conducted experiments we used the realistic leave-one-user-out cross validation (Chapter 3.2.5).

6.3.1 Baseline Systems

As in previous fusion experiments with fusion systems, we need a baseline system to compare the effect of multi-modal fusion strategies against. The best single modality performance (Table 6.3) we can achieve under our evaluation specifications serves this role well.

Single Channel Performance					
Acted			**Natural**		
	Audio	**Video**	**Audio**	**Video**	
Anger	39.0%	57.0%	59.0%	60.0%	**Positive**
Disgust	32.0%	34.0%	64.0%	50.0%	**Neutral**
Fear	43.0%	11.0%	61.0%	48.0%	**Negative**
Happiness	21.0%	82.0%			
Neutral	86.0%	72.0%			
Sadness	67.0%	59.0%			
Surprise	25.0%	22.0%			
UA	45.0%	**48.0%**	**61.0%**	53.0%	**UA**

Table 6.3: Audio and video modalities perform on an equal level in the acted DaFEx corpus, the more realistic and natural CALLAS corpus clearly shows better results on prosodic observations.

Both modalities perform on an equal level in the acted DaFEx corpus (with a slightly better video channel), the more realistic and natural CALLAS corpus clearly shows better results on prosodic observations and therefore the audio channel outperforms the facial modality (though former studies such as (Zeng *et al.* [184]) have shown a rather contrary behaviour). These results may be caused by the appliance of mood-inducing sentences for CALLAS sample generation: As the DaFEx corpus features professional actors, vocal and facial expressions are be expressed on an equally convincing level, while the inexperienced CALLAS study participants focus strongly on expressing the given sentences verbally.

6.3.2 Comprehensive Selection of Synchronous Fusion Strategies

Data in the DaFEx and CALLAS corpus is presented as segmented samples with audio and video modalities. The typical way to handle this scenario is to to consider the affective channels in a synchronous manner. Because no annotation spanning the whole recording sessions exists and therefore time gaps between the segments are not covered, model-based fusion strategies (Chapter 4.1.6) that try to model temporal relations between affective cues in modalities will not be applied within the comparison. The preconditions for these asynchronous fusion approaches will be explained comprehensively in Chapter 7.1.2. It should be noted that there is the theoretical possibility to apply asynchronous and event-driven fusion strategies within the segmented samples. This would be possible because the segments describe rather long periods of time. However, the missing time gaps between segments are nevertheless sub-optimal and we will stick to the common approach in a scenario with pre-segmented samples: We try to encode temporal information is within the features by calculating long-term statistics over extracted short-term features (Table 6.2).

In the following, we will formally describe a wide range of possible synchronous fusion approaches to be compared within the study. For the explanation of presented fusion algorithms the following annotations are used: The decision of ensemble member t for class n is denoted as $d_{t,n} \in \{0, 1\}$, with $t = 1..T$ and $n = 1..N$ and $d_{t,n} = 1$ if class ω_n is chosen, $d_{t,n} = 0$ otherwise. The support given to each class n (i.e. the calculated probability for the observed sample to belong to single classes) by classifier t is described as $s_{t,n} \in [1, 0]$.

Decision Level Fusion

Voting (Chapter 5.2.2) could be considered the most generic approach to decision level fusion, because it simply combines class labels gained from T classifiers by summing up decisions. The ensemble decision for an observed sample x is chosen to be the class ω_n which received the most votes (decisions) v_n. A definite decision is only guaranteed if an odd number of ensemble members handle a two-class problem (thus it is not capable of producing definite decisions in many practical applications and is therefore often replaced by the weighted variant).

$$v_n(x) = \sum_{t=1}^{T} d_{t,n}(x) \tag{6.1}$$

In weighted majority voting each vote is associated with a pre-calculated weight (in our case weights are determined by evaluations of classifiers on training data) of the ensemble member. Ties are not likely to happen this way, which makes the weighted variant more suited for most classification problems.

Class Label Combination							
	Acted			Natural			
	Voting	**BKS**	**CS**	**Voting**	**BKS**	**CS**	
Anger	57.0%	53.0%	35.0%	59.0%	62.0%	60.0%	**Positive**
Disgust	34.0%	45.0%	38.0%	64.0%	62.0%	63.0%	**Neutral**
Fear	11.0%	30.0%	44.0%	61.0%	56.0%	61.0%	**Negative**
Happiness	82.0%	84.0%	53.0%				
Neutral	72.0%	85.0%	90.0%				
Sadness	59.0%	51.0%	66.0%				
Surprise	22.0%	35.0%	27.0%				
UA	48.0%	**55.0%**	50.0%	**61.0%**	60.0%	**61.0%**	**UA**

Table 6.4: The BKS approach clearly outperforms the uni-modal baseline system on the acted dataset. No multi-modal effect is observable for the natural corpus.

Another way of combining the class labels generated by ensemble members is to construct a lookup table. This method is introduced by Huang and Suen [88] as behaviour knowledge space (BKS). During training the table counts combinations of labelling outputs together with the true class and occurrences of this composition. Test samples then are compared to that table and the true class for which the currently observed labelling combination was recorded most often gets chosen as ensemble decision (Chapter 5.2.2).

Table 6.4 shows results gained with the described class label combination rules. We will later on summarize, compare and interpret results of all implemented fusion strategies, but here we can already see that the BKS approach clearly outperforms the uni-modal baseline system on the acted dataset, while no multi-modal effect is observable for the non-acted corpus. Note that the BKS approach is the worst in this category for natural data.

Algebraic combiners for probabilistic outputs mathematically compute the ensemble decision from probabilities for each class over all classifiers. The Maximum Rule and Minimum Rule respectively choose the maximum or minimum support generated by T ensemble members. The ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.

$$\mu_n(x) = max, min_{t=1..T}\{s_{t,n}(x)\} \tag{6.2}$$

The Sum Rule simply sums up the support given to each class ω_n in order to generate total support μ_n for each class. By averaging the support ($\frac{1}{T}$ serves as normalization factor) given to each class ω_n we obtain the Mean Rule. When additionally adding classifier weights w_t, the Weighted Average method calculates total support μ_n for class n as:

$$\mu_n(x) = \frac{1}{T}\sum_{t=1}^{T} w_t s_{t,n}(x) \tag{6.3}$$

By multiplying the support given to each class ω_n, the Product Rule determines total support μ_n for class n as:

$$\mu_n(x) = \frac{1}{T}\prod_{t=1}^{T} s_{t,n}(x) \tag{6.4}$$

The following two combination rules make more extensive use of continuous outputs of ensemble classifiers. Given sample x, the decision profile $DP(x)$ for T ensemble members contains the probability distributions among N classes:

$$DP(x) = \begin{array}{|c|c|c|} \hline s_{1,1}(x) & ... & s_{1,N}(x) \\ \hline ... & ... & ... \\ \hline s_{T,1}(x) & ... & s_{T,N}(x) \\ \hline \end{array} \tag{6.5}$$

Decision template DT_n can then be defined for each class ω_n as respective decision profile during training, averaged by the cardinality of observed class. Given an unlabelled test-sample x, we first construct $DP(x)$ from ensemble members and then calculate similarity (as Squared Euclidean distance) S between $DP(x)$ and the decision template DT_n for each class ω_n. Finally the most similar class is chosen as ensemble decision.

Further utilisation of decision templates is based on on the Dempster-Shafer theory of evidence (Shaefer [152]). It can be applied to decision making by interpreting the classifiers outputs as a measure of evidence. Instead of similarities, proximities and resulting beliefs (evidence) is calculated. This represents the belief in one classifier

Algebraic Combination								
Acted								
	Max	**Min**	**Mean**	**Sum**	**Avg**	**Prod**	**DT**	**DS**
Anger	48.0%	44.0%	52.0%	52.0%	58.0%	50.0%	51.0%	48.0%
Disgust	31.0%	39.0%	38.0%	38.0%	41.0%	39.0%	41.0%	41.0%
Fear	22.0%	41.0%	36.0%	36.0%	28.0%	38.0%	30.0%	31.0%
Happiness	80.0%	44.0%	79.0%	79.0%	83.0%	79.0%	67.0%	67.0%
Neutral	84.0%	73.0%	78.0%	78.0%	77.0%	77.0%	81.0%	81.0%
Sadness	69.0%	59.0%	71.0%	71.0%	66.0%	70.0%	61.0%	59.0%
Surprise	16.0%	39.0%	26.0%	26.0%	23.0%	27.0%	22.0%	25.0%
UA	50.0%	48.0%	**54.0%**	**54.0%**	**54.0%**	**54.0%**	50.0%	50.0%
Natural								
	Max	**Min**	**Mean**	**Sum**	**Avg**	**Prod**	**DT**	**DS**
Positive	62.0%	56.0%	59.0%	59.0%	61.0%	59.0%	57.0%	56.0%
Neutral	55.0%	61.0%	58.0%	58.0%	58.0%	58.0%	60.0%	62.0%
Negative	64.0%	55.0%	59.0%	59.0%	58.0%	59.0%	59.0%	59.0%
UA	**60.0%**	57.0%	59.0%	59.0%	59.0%	59.0%	59.0%	59.0%

Table 6.5: On acted data, one group of combination rules (Mean, Sum, Avg and Prod) establishes a recognizable multi-modal effect, but their equal performance leads to an impression of interchangeability. This impression is solidified when looking at results of fusion with natural data, where more or less all approaches lead to similar results.

correctly classifying observed instance into respective classes. Following Dempsters rule of combination, these beliefs can be multiplied throughout the ensemble in order to obtain the final decision.

Table 6.5 describes results of algebraic combination rules for the probabilistic outputs of the modality classifiers. On the acted DaFEx corpus we identify a group of rules that do not lead to a significant multi-modal effect and a second group that does. Within this second group (Mean, Sum, Avg and Prod) however, we get an impression of interchangeability as they all perform on equal level. This impression is solidified when looking at results of fusion with natural data, where more or less all approaches lead to similar results (that still produce no multi-modal effect).

Feature, Hybrid and Score Level Fusion

The feature and score level fusion have been discussed in detail in Chapters 4.1.2 and 4.1.4. In this study we use the term hybrid fusion to characterise fusion techniques that incorporate classifiers with merged features into used ensembles and therefore combine decision and feature level fusion. Of course this approach is applicable to most ensemble combination rules discussed so far, but we decided to develop a refined fusion scheme with two variants in order to explore the capabilities of hybrid fusion.

Feature, Hybrid and Score Level Fusion					
Acted					
	Feature	**OvR**	**OvR-Specialists**	**Stacking**	**Grading**
Anger	54.0%	53.0%	59.0%	53.0%	60.0%
Disgust	36.0%	34.0%	31.0%	40.0%	44.0%
Fear	36.0%	36.0%	40.0%	39.0%	18.0%
Happiness	79.0%	83.0%	82.0%	72.0%	80.0%
Neutral	77.0%	79.0%	76.0%	74.0%	89.0%
Sadness	70.0%	71.0%	70.0%	61.0%	64.0%
Surprise	26.0%	25.0%	21.0%	28.0%	23.0%
UA	54.0%	**55.0%**	54.0%	52.0%	**54.0%**
Natural					
	Feature	**OvR**	**OvR-Specialists**	**Stacking**	**Grading**
Positive	57.0%	59.0%	60.0%	59.0%	67.0%
Neutral	59.0%	59.0%	58.0%	57.0%	50.0%
Negative	62.0%	60.0%	63.0%	64.0%	49.0%
UA	59.0%	59.0%	**60.0%**	**60.0%**	55.0%

Table 6.6: Results of feature, hybrid and score level approaches confirm the so far established picture of interchangeability, as performance on acted data and natural data lies in very close respective ranges. If one fusion strategy seems superior on one corpus, it turns out to be inferior on the other.

The One Versus Rest (OvR) approach trains N classifiers on every available feature-set (excluding merged features), each specialised in recognising one of N classes. This breakdown on several two-class classification problems is done by relabelling. Additionally the ensemble is completed by one multi-class classification model trained on the merged feature set. Given test-sample x, variant one multiplies probabilities gained from classifiers trained on recognising class n with the associated probability generated by the multi-class classification model. This is done for classes $1..N$ and the class with

the highest support gets chosen as ensemble decision. Variant two (OvR-Specialists) chooses among the two-class classification models the most promising one for every class. The specialist's probability is then summed with the respective probability from the multi-class classification model.

Results of feature, hybrid and score level approaches (Table 6.6) only confirm the so far established picture of interchangeability: Performance on acted data and natural data lies in very close respective ranges. If one fusion strategy seems superior on one corpus, it turns out to be inferior on the other.

Summary of Results

Results for DaFEx (acted) and CALLAS (naturalistic) corpora are summarised in Table 6.7. For the acted data we observe an impressive improvement of up to 7% multi-modal effect compared to classification results of the single expert system (video modality). Here, decision level fusion with class label combination in lookup tables (BKS) and a custom hybrid level approach (OvR) turn out to give the best performance. However, apart from some outliers (e.g. minimum rule) all applied fusion strategies operate on a similar level of quality. In case of the more natural CALLAS corpus no improvement is achieved compared to the single expert system (audio modality). Again, results range in a very tight interval, but the multi-modal effect is not existent or even negative. These outcomes go in line with insights that were later published by D'Mello and Kory [48], stating that a greater multi-modal effect can be expected in acted settings.

Across corpora, simple fusion techniques like feature fusion as well as mean, sum or product rule perform on a stable basis. More elaborate strategies seem to be more reliant on the structure of observed data. For example, the CS method (Chapter 5.1.3) generates the desired flattening effect among classes on the CALLAS corpus and therefore lists among the best fusion approaches. In contrary, needed specialist selection seems to be harder on the DaFEx corpus and it ranges among worst combination rules. Sophisticated ensemble strategies bare the potential to outperform more simple ones, but success is not guaranteed.

All in all, differences in accuracy tend to be rather small among all considered fusion techniques. Table 6.7 clearly shows that fusion approaches perform very similar on the same corpus and if one is slightly superior on one kind of data, it is not on the other. From this point of view we can affirm the impression that the *No Free Lunch* theorem (Section 6.1) holds for the various synchronous fusion approaches we have included within the comparison study.

Summary of Results			
Approach	**Algorithm**	**Result**	**Effect**
DaFEx (Acted)			
Baseline	Single Expert (Video)	48.0%	**-**
Class Label Combination	BKS	55.0%	**+ 7.0%**
Algebraic Combination	Mean / Sum / Avg / Prod	54.0%	**+ 6.0%**
Feature and Hybrid Level	OvR	55.0%	**+ 7.0%**
Score Level	Grading	54.0%	**+ 6.0%**
CALLAS (Natural)			
Baseline	Single Expert (Audio)	61.0%	**-**
Class Label Combination	Voting / CS	61.0%	**-**
Algebraic Combination	Max	60.0%	**- 1.0%**
Feature and Hybrid Level	OvR-Specialists	60.0%	**- 1.0%**
Score Level	Stacking	60.0%	**- 1.0%**

Table 6.7: The insights of later surveys, stating that a greater multi-modal effect can be expected in acted settings, is validated be these results. All in all, fusion approaches perform very similar on the same corpus and if one is slightly superior on one kind of data, it is inferior on the other.

6.3.3 The Barcode Pattern

In Figure 6.2 recognition results are visualized per sample, as we compare the prediction for each sample with its real label. If the sample was correctly classified, it is marked with a white square, otherwise with a black one. Each column represents one sample of the data set and each row stands for the used fusion method. The first row, for instance, visualizes classification results obtained for the single audio channel. The acted DaFEx corpus is shown on top, the natural CALLAS corpus on the bottom. We can for example infer from the DaFEx pattern on top of Figure 6.2 that the first two samples were correctly classified by audio, video and most fusion schemes, while sample three and four were obviously misclassified.

A characteristic shown by this visualisation on both corpora is the behaviour of fusion schemes in relation to single modalities. Depending on the outcomes of audio and video, there is a clear trend of forming white and black vertical columns within the picture: If both modalities classify correctly, most fusion approaches do so too; if both channels misinterpret the sample, most fusion strategies fail. Single coloured columns show the

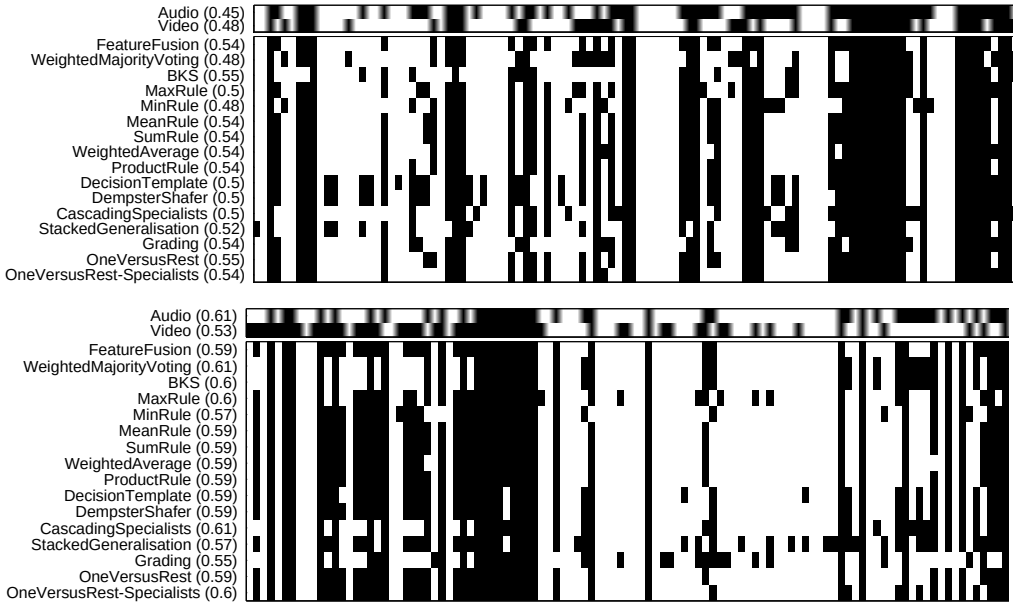


Figure 6.2: Within this visualisation, recognition results are visualized per sample. Each column represents one sample of the data set and each row stands for the used fusion method. If a sample is correctly classified, it is marked with a white square, otherwise with a black one. Characteristics of fusion methods can consequently be inferred in a sample-by-sample manner. The acted DaFEx corpus is shown on top, the natural CALLAS corpus on the bottom.

tendency of fusion methods to gain and lose overall accuracy on the same samples, we call this the barcode pattern. Especially algebraic combiners like the sum or product rule amplify consistent (correct or incorrect) ensemble decisions because of their inherent combination rules. The barcode pattern can be interpreted as hint to the *No Free Lunch* theorem.

An exception to this trend is shown by some decision level approaches (BKS, decision template and dempster shafer) and the meta-learners of stacked generalisation and grading: These approaches are meant to learn the behaviour of available modalities. The logical connection between observed ensemble members' decisions and the actual true-class enables the phenomenon of error-learning and therefore the potential of generating correct predictions though both modalities classify incorrect. This is visualized by white break-ups in black columns caused by consistent miss-classification of audio and video. Figure 6.2 unfortunately show that this desirable error-learning is also the reason for these ensemble methods to predict wrong classes even though both modalities chose the correct one – an undesirable characteristic that is not likely to be manifested by more simple fusion schemes.

6.4 Conclusions

In this chapter we performed a comprehensive comparison of synchronous fusion techniques for multi-modal affect recognition tasks. Our experiments were run on two Italian emotion corpora featuring vocal and facial modalities: The acted DaFex corpus and the more natural CALLAS expressivity corpus. Implemented fusion rules include feature, decision, score and hybrid level strategies. First off, results are in line with insights gained by other studies that state greater multi-modal effects in acted settings. In fact, some fusion approaches even resulted in negative effects on natural data. Looking at each corpus in isolation, differences in recognition performance for the included fusion schemes seem mostly neglectable. We find that multi-modal fusion is in almost any case at least on par with single channel classification, though homogeneous results within corpora point to interchangeability between concrete fusion schemes. If we however identify the best among these and look at their performance on the respective other corpus, we see that their application is not advisable in the new case. If one fusion strategy seems superior on one corpus, it turns out to be inferior on the other. A general advise is not possible, affirming our hypothesis that the *No Free Lunch* theorem is valid for the various synchronous fusion approaches.

Indications for the interchangeability of these fusion schemes are given by barcode pattern visible in our visualisation (Section 6.3.3). Because of segmented samples and the single high level statistical feature set for each modality, there are only to synchronized sources of information for the fusion process available. Trends show that whenever both modalities classify correctly, then the fusion approach succeeds. If both channels misinterpret the sample, the fused result is also often incorrect. Of course there are sophisticated methods such as the appliance of weights or lookup tables to be included in the fusion process to counteract and at least generate a correct result if one modality points to the right decision. However, the root of this effect may lie in the synchronous treatment of modalities given by the segmentation of the samples. The boundaries of a multi-modal sample are mostly dictated by a prominent modality (e.g. a spoken word or phrase) and forced upon remaining modalities. The synchronous fusion rules consequently have to deal with channel information that may be suboptimal, as the affective cues in other channels may occur time shifted and may be cut off or scattered sparsely within the common segmentation bounds. We refer to this phenomenon as the *segmentation problem*. In the following chapter we will therefore analyse this problem in detail and try to solve it with asynchronous and event-driven treatment of considered modalities, that are analysed in the shortest frames allowed by selected feature extraction approaches.

Chapter 7

Asynchronous and Event-driven Approaches to Model-based Fusion

Introductory experiments in Chapter 5 and the subsequent systematic comparison of fusion strategies presented in Chapter 6 investigated the combination of multi-modal information on the feature, decision, score and hybrid level. In most cases we were able to surpass the uni-modal baseline system. However, on non-acted natural affective depictions this was not always the case and closer inspection of gained results gives the impression of interchangeability of investigated synchronous fusion strategies. We have however by now not yet exploited the whole range of possible approaches to modern emotion recognition, as all these fusion methods are limited by following a synchronous strategy. Relevant affective cues are expected to appear in parallel time segments accross considered modalities. Therfore these methods are not accounting for temporal relations between modalities. Beyond these synchronous approaches we are able to find fusion models that implement asynchronous treatment of affective channels: D'Mello and Kory [49] subsume fusion strategies that aim to model asynchronous temporal relations and consider differing temporal flows of affective manifestations in affective channels under the term model-based fusion (Chapter 4). In this chapter we will investigate these approaches in detail and try to contribute to this asynchronous fusion approach with the concept of affective events and their use in affect recognition. Instead of designing a fusion algorithm that asks for input from modalities at a fixed rate, in (Lingenfelser *et al.* [111]) we model a fusion system that is able to register and remember affective cues whenever they occur. These short timed events are recognized in each modality separately. They can on the one hand directly fit the sought emotional category but on the other hand presented algorithms will also be able to handle an additional abstraction

layer in which the events do not necessarily share the label of the target class. In this case these indirect events indicate the abstract target class with increased occurrences over time.

7.1 Solving the Segmentation Problem

First of all there is a certain classification setting in which asynchronous fusion is unlikely to be applicable. In preceding chapters we evaluated data in samples. This means that segments containing emotional expressions were cut out of whole recording sessions and portions of data not contained in the segmentations are not of interest for the evaluation (Figure 7.1).

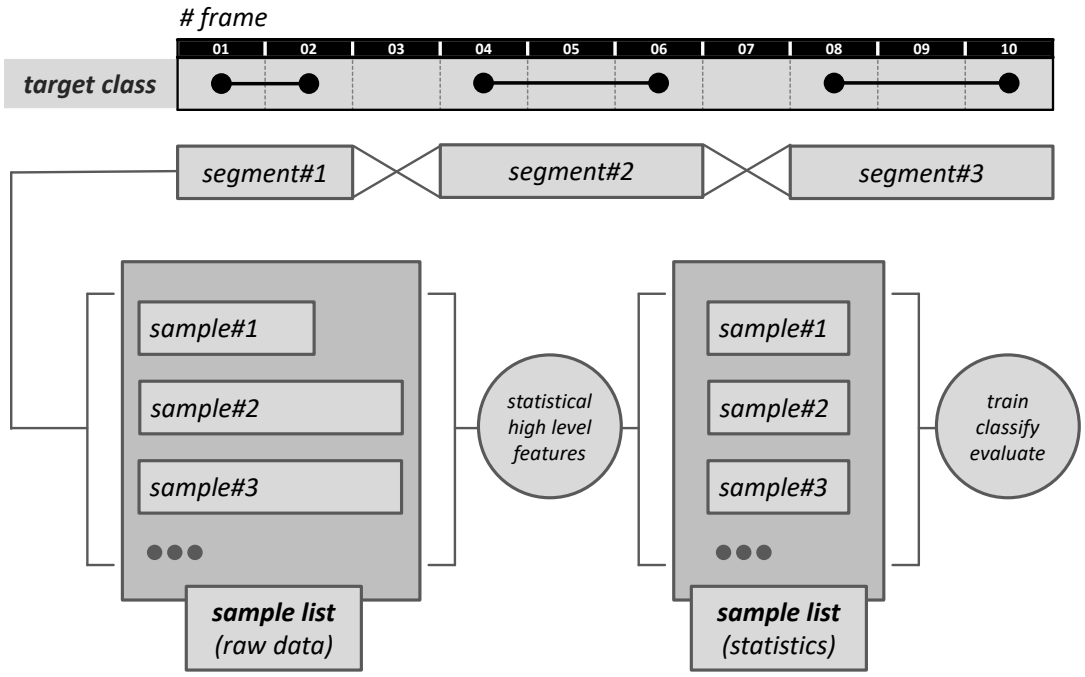


Figure 7.1: The traditional segmentation-based sample generation procedure: An annotation track for a given target class defines data segments of interest. As these samples can be of variable length statistical high level features are calculated over the raw data segments. Resulting samples are stored in a sample list that is used to train and evaluate the recognition system.

Extracted samples of raw data are processed for classification mostly via the extraction of statistical high level features (Section 3.1.1) that compactly describe several frames

of recording. The resulting collection of samples is stored in a sample list that is used for the training / classification / evaluation process. Given this setting plus samples that contain high level features for more than one modality we are more or less bound to synchronous fusion approaches: We either merge the statistical features of each modality (feature level fusion) or classify each modality separately and merge the results on any level described in Chapter 4. However, the temporal information within and / or across modalities is at best encoded in the high level features, but impossibly in the synchronous order in which we consider input from modalities.

So let us have a closer look at the temporal phenomena that may actually cause the synchronicity problem for common fusion strategies. There is evidence from psychological studies that the temporal dynamics of emotional displays is not necessarily the same across all modalities. To illustrate this, we exemplarily inspect a typical behavioral pattern expressing embarrassment. According to Keltner [93], the display of embarrassment usually starts with a gaze down followed by a sequence of smiles, gaze and head shifts. This sequence of gestures and facial expressions is maybe accompanied by some spoken words containing paralinguistic indications. It is the sequence of coherently integrated modalities that distinguishes embarrassment from related affective states, such as amusement or enjoyment. While the single modalities are correlated to each other, they seldom start and end exactly at the same point in time, but follow each other with a small time lag or partially overlap in time (Figure 7.2).

This means problems for classical fusion algorithms that are applied in the segmentation-based scenario: The on and offset of a relevant time-interval in one modality is used to call fusion techniques for classification throughout all available modalities, e.g. the voice activity in the audio channel is used to detect a spoken word, phrase or sentence and during this time other modalities such as eye gaze, head orientation, facial expression etc. are also considered (Figure 7.2). Consequently, the segmentation of an expressive cue in one modality is forced upon other available channels. What if no observable effects are happening in other considered modalities at this point in time, as emotional reactions are time-shifted between modalities or not present at all? Meaningful information in additional modalities is assumed - but it is not guaranteed. A fusion strategy that expects usable information in all modalities at a certain point in time will often fail in such situations. Thus, synchronously cutting segments through multi-layered signals does seem to be undesirable. This approach may be sufficient in scenarios with acted or very clearly depicted emotions in which the affective cues emerge very coherently in all recorded channels, but it may fail in non-acted and more subtle situations in which affect is shown only in some of the channels or in a very unaligned and incoherent way.

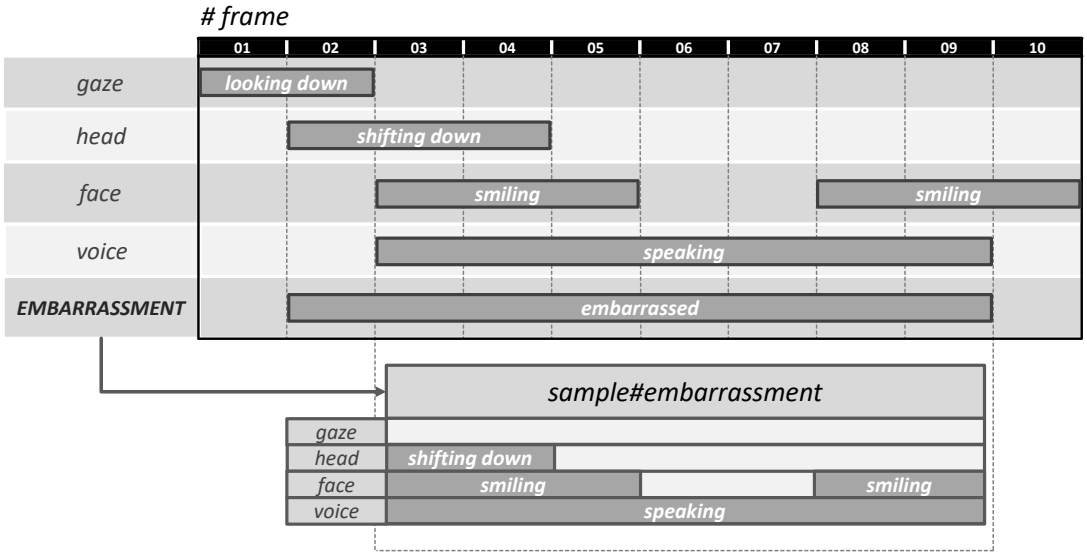


Figure 7.2: The segmentation of the given target class (e.g. embarrassment) is here exemplarily given by the voice activity detection in the audio channel. During voice activity other modalities such as eye gaze, head orientation, facial expression are considered in addition to the vocal modality. We can observe that in this example there are only sparse cues from the head orientation included in the resulting sample. The embarrassment indicating gaze is completely left out as it occurred in preceding frames.

So how can the problem of non-aligned cues in multiple signals be solved in order to recognize emotions (especially natural ones) better?

7.1.1 Framewise Classification

A first step towards the asynchronous fusion approaches we are aiming to investigate in this chapter is to narrow down the analysis windows for the concerned modalities. The segmentation-based approach described in this section tailors a sample's length based on annotation or some kind of detected activity. Another way to treat a signal to be classified is to observe it as a sequence of small time slices - the so called frames. The size of these frames is typically given by the smallest chunk possible in order to describe it's content with features. Given e.g. a video stream, a single recorded image would be sufficient while in case of audio we could consider calculating short-term features like MFCCs (Section 3.1.1), which are typically calculated over 25 milliseconds of data at least. Unlike in the case of segmented samples of variable length, we could now exclude the process of forming statistics over these short-termed features, as every calculation frame is of exactly the same length (Figure 7.3).

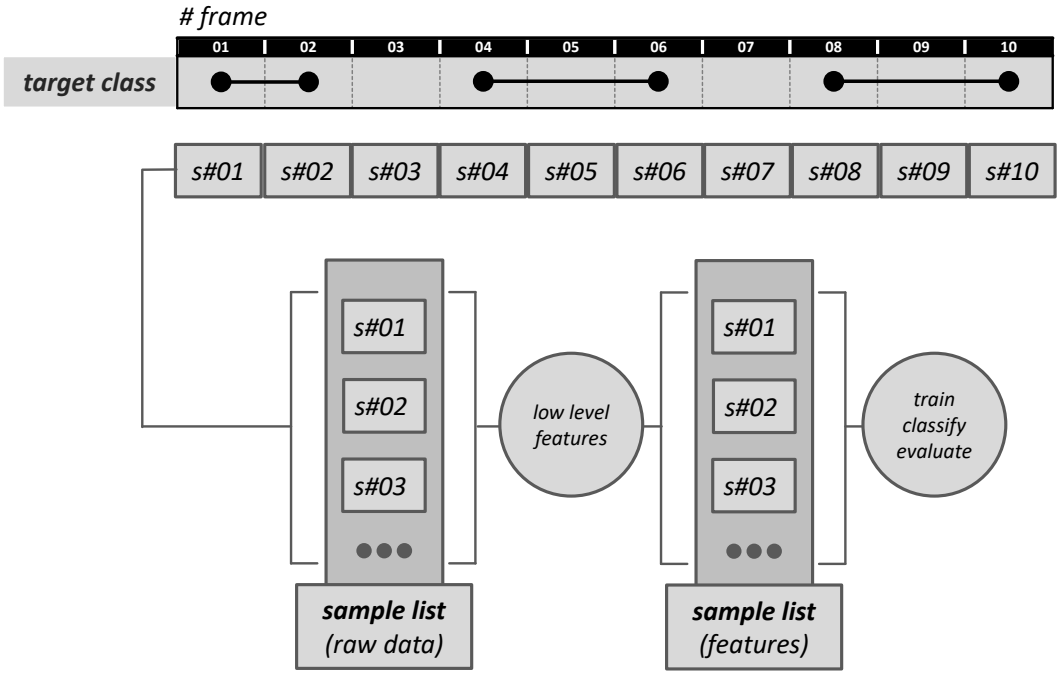


Figure 7.3: In a framewise classification scenario a signal stream is treated as a sequence of short termed frames. The size of these frames is typically given by the smallest chunk possible in order to describe it's content with low level features.

The advantage of a framewise recognition approach is that it enables us to continuously classify whole streams of data. This is especially preferable in practical applications as it enables an affect recognition system to classify input in real-time (or near real-time respectively - given the size of the frames). Furthermore this framewise treatment of signal streams opens up a whole new possibility: As the frames are defined to occur in an ordinal sequence and describe the signals without cut-out segments, past frames can be included in the process. Classification algorithms such as recurrent neural networks (Section 3.2.4) that implement memory capabilities can use the information included in several successive frames to model the temporal flow within a signal. This behaviour is theoretically superior to the strategy of describing the temporal component with statistics over features.

These benefits are not obtainable without effort. As indicated in Figure 7.3, every single frame of the signal will be included in the final sample list. This means that start-to-finish annotations of whole sessions are required to give labels to each and every entry.

The annotation effort potentially multiplies in comparison to the effort to only label selected segments of sessions. Furthermore the ordinal sequence of samples / frames is demanded not only in the final real-time application (where it should occur naturally) but also at training and evaluation stage. If we are however able to guarantee framewise classification of input modalities we also have the possibility to implement asynchronous fusion systems.

7.1.2 Synchronous versus Asynchronous Fusion Systems

We will from now on in this chapter assume the preconditions described in the previous section as given:

- Framewise classification (as indicated in Figure 7.3).
- Start-to-finish annotations of whole sessions (in contrast to sample lists with cut-out segments).
- Ordinal sequence of samples (frames) not only in real-time application but also at training and evaluation stage.

Consequently we can debate two alternatives to handle the incoming multi-modal frames for fusion - a synchronous or asynchronous strategy. Figure 7.4 shows the schematic approach to synchronous handling of multi-modal frames. A synchronous fusion approach can be defined by the consideration of multiple modalities within the same time frame. This task is straightforward if frames are of same size for all modalities, otherwise one would need to manipulate the data with e.g. delayed decisions until the longest frame has been processed or statistics over features of shorter frames to fit the longest one.

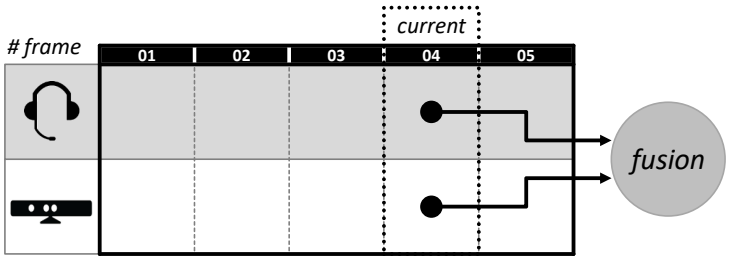


Figure 7.4: Synchronous fusion approaches are characterized by the consideration of multiple modalities within the same time frame.

If we consider the samples resulting from a segmentation-based approach as (arguably rather long) frames, all fusion strategies applied in our so far discussed experiments can be regarded as synchronous. In addition to these, there is a fair amount of studies that incorporate synchronous fusion techniques for combination of observed signals, but the success of fusion is obviously not primarily dependent on the chosen algorithm (though of course there are differences in performance between the single fusion strategies, Chapter 6). In a meta review of multi-modal affect recognition systems, D'Mello and Kory [48] point out, that improvements based on integrating multi-modal information are far more likely to be achieved on acted data than on natural or semi-natural recordings. As we have already hypothesized in Section 4.2, this may be due to the assumption that in an acted and exaggerated scenario affective cues appear more consistent across modalities than in a natural depiction of emotion.

The question at this point is how we can use the framewise approach we have defined in order to enhance the multi-modal fusion approach? We have seen by now that we have to reject the assumption that all relevant cues happen in all modalities at the same time, which is an implicit assumption of synchronous fusion approaches (Figure 7.4). In Figure 7.2 we can easily identify single frames in which not all modalities are including target class inducing information. To relief this problem information from surrounding frames needs to be included (Figure 7.5). In practice this means mostly the consideration of preceeeding frames, as in most real-time applications the observation of succeeding frames leads to delays in the whole recognition pipeline. An early adoption of this requirement is presented by Mansoorizadeh and Charkari [114]. The study describes natural differences in temporal patterns of multiple modalities and states that affective cues may occur temporal unaligned in different signals. An interesting solution is presented in which the fusion of modalities happens on feature level by constructing a unified feature vector out of the individual (in the case of the evaluation study) voice and face channels. The feature space is not updated synchronously for all modalities, but asynchronously with differing sampling rates. This way, the classification model associated with the merged feature space is working on information, that is (in respective parts) updated on basis of decisions of the single modalities, not at a unified, synchronous rate dictated by the system.

Asynchronous fusion approaches on the feature level are, however, seldom. A more common way to relate temporally unaligned cues for proper fusion is the use of dynamic classifiers such as hidden Markov models. Dynamic classifiers work on streams of dynamic length and fuse modalities without the need to force decisions from all channels in every timeslice. They share the ability to model temporal relations between the streams

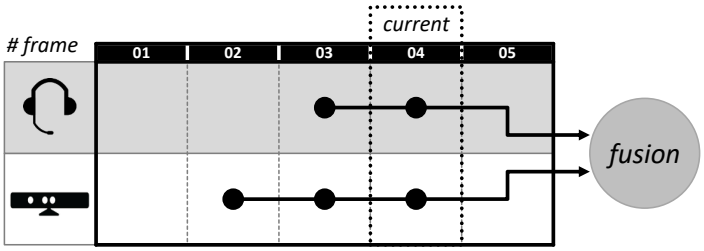


Figure 7.5: Asynchronous fusion approaches distinguish themselves from synchronous combination strategies by referring to past time frames - often with some kind of memory support. Therefore they are able to consider temporal shifts between affective depictions in modalities.

by learning when to combine multi-modal information. Song *et al.* [158] however state that the audible cues occur delayed from visible hints and apply coupled hidden Markov models (Brand *et al.* [19]). Like multi-stream hidden Markov models these coupled models consist of multiple hidden Markov models (one for each stream of data). Discrete (hidden) nodes of each included model at time t are conditioned by discrete nodes of all other models within the system at time $t-1$. This way the coupled hidden Markov model is able to learn intrinsic temporal correlations between modalities. Song *et al.* [158] use this approach to integrate three affective channels for emotion recognition based on facial expressions, visual speech (feature extraction of mouth and lip regions) and low level audio features. Nefian *et al.* [125] use coupled hidden Markov models to fuse visual speech and Mel frequency cepstral coefficients for speech recognition tasks. Zeng *et al.* [186] applied multi-stream fused hidden Markov models to recognize cognitive states and prototypical emotions. These fused multi-stream models are further developed with respect to the connections between hidden nodes. Instead of coupling each and every hidden node in the system, the single hidden Markov systems are first trained independently, and in a second step the number and parameters of the node connections are determined (Pan *et al.* [130]). One of the expected advantages is the robustness of the resulting fusion system because if one model is poorly trained (e.g. due to noisy data), the system as a whole is still capable of correct classification as the single components are first trained in isolation and the connections are added afterwards. Most of these multi-modal Markov models share the drawback of computational complexity. In order to mitigate this problem Wöllmer *et al.* [178] suggest a multi-dimensional dynamic time warping algorithm for hybrid fusion of asynchronous data streams, which couples low decoding time with high the high flexibility of the asynchronous hidden Markov approach.

Other ways of applying asynchronous fusion to multi-modal data streams can be found in

the form of recurrent neural networks with memory capabilities (Brückner and Schuller [21], Hochreiter and Schmidhuber [83]). They are able to learn a history of past frames in several modalities and take them into account for classification, in particular in the form of long short-term memory neural networks. These replace common network nodes with memory cells. These cells give the network the possibility to learn when and how long to remember past information (Wöllmer *et al.* [179], Chapter 3.2.4).

7.2 Affective Events

All the asynchronous, model-based fusion approaches we have seen so far share an internal structure: They work directly on the incoming data streams and resulting features. The asynchronous nature of observed modalities is mainly factored in by back-facing connections between nodes within the models. The proposed concept of affective events works a bit different. The data streams are decoupled from the final fusion system, instead additional recognition components search for affective cues in the observed modalities and broadcast their findings. The fusion algorithm itself works as a client that processes the registered events with respect to the time of their occurrence (Figure 7.6). This architecture fully respects the asynchronous nature of affective channels and is conceptually built to handle differing characteristics of respective data streams (sample rate, feature range, etc.).

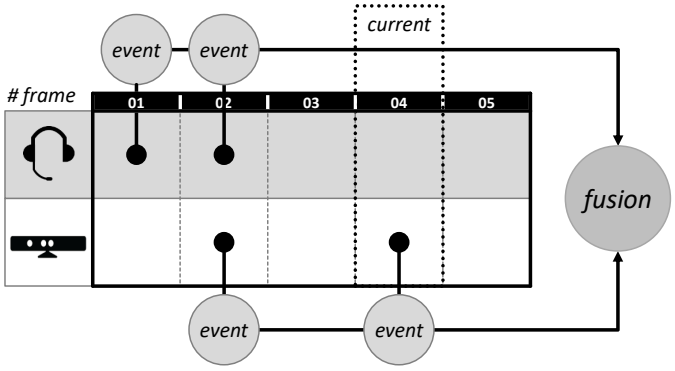


Figure 7.6: Event-driven Fusion Scheme. The target class is not directly classified, but target class indicating events are recognized by accordingly trained models. The final classification has to be algorithmically derived from found events.

Events can hereby be defined as manifestations of the sought affective target class that are emitted by the involved modalities and occur asynchronously across channels. They are per definition strongly coupled to some kind of activity detection in the respective

affective channels. The assurance of meaningful activity can be as simple as guaranteeing stable tracking in frames containing video data of e.g. faces for facial analysis or hands for gesture analysis. More elaborate activity measures include signal processing steps such as the calculation of a signal-to-noise ratio to determine parts of the audio channel that probably hold paralinguistic information or peak detection in a galvanic skin response signal to differentiate the baseline from sequences of affective stimulation. Frames that do not pass an activity check will not be qualified to generate affective events.

7.2.1 Events as an Additional Abstraction Layer

"Most approaches to emotion recognition are based on categorical emotion theories (Ekman and Friesen [57]), which model emotions as distinct categories, such as joy, anger, surprise, fear or sadness, or dimensional models, which characterize emotions in terms of several continuous dimensions" (Mehrabian [120]), such as pleasure, arousal and dominance (Chapter 2). Classifiers are usually trained to map relevant features directly onto these discrete emotion categories or a continuous multidimensional space. So typically the probabilistic outputs of modality-specific classifiers relate to the target class of the recognition system. They are combined by a fusion strategy and an agreeing decision on the sought class is determined among the considered modalities.

At this point we want to examine an alternative to these direct emotion recognition strategies. In their approach to multi-modal fusion Mansoorizadeh and Charkari [114] use the term *event* to describe temporally unaligned manifestations of an observed emotion throughout considered modalities. This emphasizes the asynchronous nature they affiliate with the expression. We want to take this definition a step further and consider that these asynchronous manifestations do not need to directly equate to the overall affect. Instead of directly recognizing affective states from relevant features, we search for indicative events in available modalities that can be algorithmically interpreted for target class estimation. This means we introduce events as an abstraction layer. An intermediate layer of representation has also been suggested by Mortillaro *et al.* [122] for emotion recognition tasks. They propose to use expressive features for assessing appraisals, such as subjective pleasantness, which in turn could be employed for assessing emotional labels. Mortillaro *et al.* argue that the introduction of an additional layer could contribute to a higher level of interpretability of machine learning results. Even though our events do not correspond to appraisals, they represent meaningful interpretation units that lie between expressive features and emotional labels.

7.2.2 A Practical Example

To explain the event-driven approach further, we will in the following sections use the scenario of recognizing a speaker's level of enjoyment in a natural conversation as a case example (which will be further motivated and utilized for evaluation in Chapter 8). The goal is to not recognize enjoyment directly but to derive it from occurrences of smile and laugh events that can be related to a significant level of enjoyment. In many databases, there is only one annotation for all recorded modalities. This annotation defines segments of relevant information that are considered during training and evaluation in all affective channels. This approach supports the - potentially wrong - premise of synchronous fusion approaches that assumes that all channels contain meaningful information and relevant affective effects within the segment. This is exactly the assumption model-based, asynchronous and especially event-driven fusion systems try to avoid. While asynchronous models such as neural networks may be able to overcome the problem by dynamically deciding how many past frames should be considered in each modality, we need uni-modal annotation tracks to segment the individual events.

Figure 7.7 depicts a multi-track annotation that shows how enjoyment is characterized by a repetitive sequence of voiced laughs and visual smiles. The figure gives an impression of which events to expect when observing enjoyment. Even though laughs and smiles are correlated, they do not start and end at the same point in time. Also there are phases of enjoyment where neither a voiced laugh nor a visual smile occurs. Hereby it is not even of highest importance whether the uni-modal event segments are labelled as indicator events or the target class enjoyment: To a greater degree this annotation clarifies the need to inspect the observed channels in isolation, as this depiction clearly shows the asynchronicity of multi-modal cues and enables a better understanding of the affective state to be recognized - for the annotator as well as for the classification systems to be trained.

To theorize this in detail let us again have a look at Figure 7.7. It shows an exemplary annotation of a full enjoyment episode aligned with various voiced and visual cues emitted by the user. For each frame a decision has to be made by the fusion system. In a synchronised fusion approach each frame is seen in isolation, i.e. a decision is derived from the multi-modal information within the frame. However, we can see that the single cues only partly overlap with the enjoyment episode. While other frames align with cues from a single modality (see e.g. frame 09), some of the frames, which are spanned by the enjoyment episode do actually not overlap with any observable cues (see e.g. frame 04). Those frames are likely to be misclassified by a synchronous fusion approach. Obviously, asynchronous fusion approaches, which take the temporal asynchronicity of

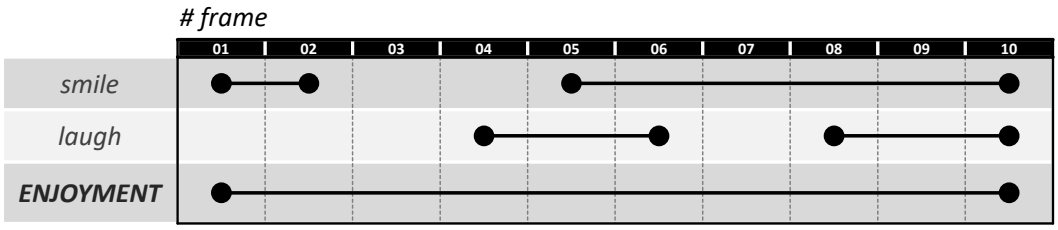


Figure 7.7: Exemplary annotations of enjoyment, voiced laughs and visual smiles. Dotted lines depict time frames in which decisions have to be made by the fusion system. Asynchronous and even-based fusion approaches have the opportunity to overcome segments with a sparse distribution of actual cues of enjoyment.

the modalities into account, should be able to catch the characteristics of the analysed data more precise. Indirect recognition approaches that use event recognition for enjoyment classification are probably able to overcome frames with sparse cues of enjoyment in current and preceding frames.

7.3 Event-driven Fusion Strategies

We have by now rejected the assumption that all relevant affective cues happen at the same time in all modalities. Therefore asynchronous events that indicate the target class in different modalities are detected. The events we have defined introduce a way to asynchronously treat modalities and establish an abstraction layer between the multimodal signal streams and the emotional label of the target class. We will now examine preconditions and concrete implementations of fusion algorithms able to handle these affective events. A real-time event-driven fusion scheme, meant to reduce negative effects of the segmentation problem (Section7.1), must meet certain requirements. The following list formulates key features an algorithm has to offer in order to be considered an event-driven fusion approach.

- **Temporal Flow**
  Once recognized, an event enters the fusion process and influences the continuous result with potency given by the strength of the recognized cue. An event's influence then has to diminish - as the moment of occurrence shifts further back in time - until the event is discarded. This way, current events are given a stronger impact on the fusion process than the ones that lie further down the time axis.

- **Reinforcement and Attenuation**
  The additional effort of annotating and detecting events in every single modality leads to a fusion model that should link independent signal events. If complementary events are detected in multiple signals during overlapping time-segments, the cues reinforce each other by amplifying the prediction probability of the continuous fusion output. On the other hand, the detection of contradictory cues leads to events that neutralise each other and therefore have a lesser negative and attenuating effect on the fusion result. This way, additional information from multiple modalities is more likely to enhance the overall classification performance.

- **Real-Time Fusion Result**
  With their reduced complexity and modular expandability event-driven fusion strategies are good candidates for practical applications. Since these situations mostly demand near real-time recognition tasks on live input, we demand all algorithms to be suitable for this task. The result of the fusion scheme is calculated by temporal influences (expressed through momentary weights) of registered events and a current fusion result has to be available at any given point in time, guaranteeing access to the latest affective estimation for all other components of the system. This circumstance is especially valuable in real-time scenarios, where reactions to changing conditions have to be carried out as fast as possible.

An early and very interesting predictor for affective states based on verbal and non-verbal behavioural events (e.g. laughs, smiles, head shakes) was proposed by Eyben *et al.* [61]. The fusion system treats events as ”words” and applies fusion on the so called ”string” level. A string is computed by joining all events (words) contained within a conversational turn of a user. The strings are converted into a feature vector using a binary bag-of-words approach (i.e. mark presence or absence of an event). The approach is borrowed from natural language processing, where words in sentences are listed (without grammar, word order or number of occurrences) to form feature vectors for further processing. This way however the feature vector contains only information whether a type of event is present in a segment or not, it does not give insight in the frequency of occurrence. Consequently the probability of the target class or amplitude within a dimensional model cannot be raised by repeatedly registered cues. Also a segmentation is given by the audio modality, which contradicts the assumptions we have made in the preceding sections. Another affect recognition approach using affective events is introduced by Gilroy *et al.* [70]. They describe an artistic augmented reality installation in which a user’s affective state is derived by events from various affective channels. The real-time application uses a fusion system, which constantly represents the

current user emotion with a vector in a dimensional emotion model. The position of this multi-modal fusion vector is updated on basis of the latest single modality contributions. This real-time approach with asynchronous updates from single modality recognizers matches our requirements well and the following algorithms build upon this fusion idea.

7.3.1 Vector Fusion

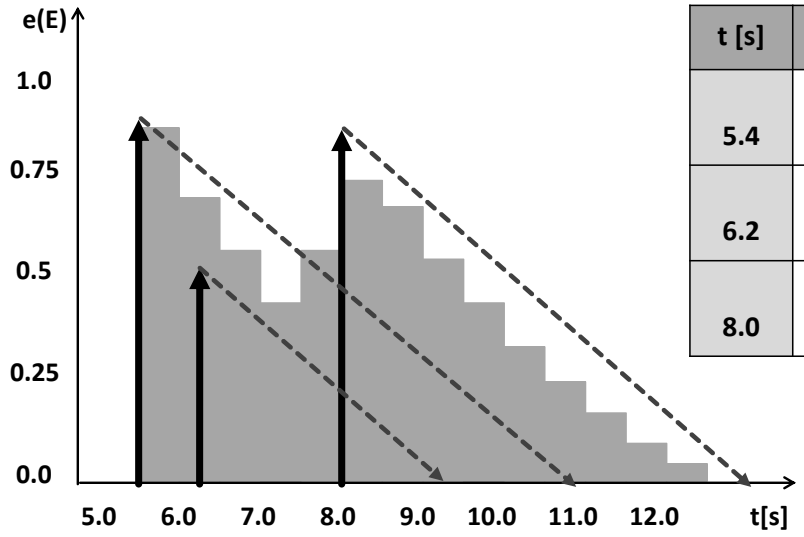
As just implied, the first proposed event-driven fusion algorithm is based on preceding work done by Gilroy *et al.* [71], which represents emotions as a vector within a dimensional emotion model. We generalize this approach by designing a fusion scheme that operates in a user-defined vector space. In the simplest scenario, the vector space is a one-dimensional axis, typically describing a likelihood between zero and one. Events, generated from observed signals, are mapped into this space as vectors. The vectors are provided with several parameters: A score is defined for each axis in the event space. This defines the position of the vector within the dimensional model. This value can be directly given by the continuous output of a classification model or can be dynamically calculated from the normed probabilities of a recognized cue, resulting in values that typically range between zero and one. Every vector is given a weight parameter, which serves as a quantifier for its impact on the calculation of the fusion result. It is defined by the modality the event is recognized in and serves as a regulation instrument for emphasizing more reliable information sources. Finally, the decay speed parameter describes the average lifespan of cues extracted from the respective signal. It is also defined for each modality and determines the time it takes for the event's influence to decrease to a neutral state and get discarded. Events that strongly indicate the target class can be given longer decay times in order to prolong their influence on the result.

At each time frame, active event vectors $e = 1 \dots E$ are decayed by multiplying each vector element with a decay factor that is calculated based on the defined decay speed, expired lifetime and the initial norm of the vector:

$$decay_e = norm_e - (lifetime_e * speed_e) \tag{7.1}$$

If the resulting norm of the decayed vector stays above zero, it remains active - otherwise the vector is discarded. Afterwards the fusion point within the vector space is calculated from all active event vectors: For each dimension $d = 1 \dots D$ of the vector space respective scores of active event vectors $e = 1 \dots E$ (modified by their weight factor) are summed up.

$$fusion\ point(d_{1 \dots D}) = \sum_{e=1}^{E} (event\ vector_{d,e} * weight_e) \tag{7.2}$$



t [s]	event	e(E)
5.4	*laugh (audio)*	0.9
6.2	*smile (video)*	0.5
8.0	*laugh (audio)*	0.8

Figure 7.8: Example schematic of the vector fusion algorithm. Three enjoyment indicating events from audio and video modality (black arrows) are successively mapped into the vector space. Their lengths decrease over time (dotted lines), therefore the fusion point moves over time with the decaying vectors.

The result is normalized by the sum of weights of all contributing event vectors.

$$fusion\,point(d_{1...D}) = mass(d_{1...D}) / \sum_{e=1}^{E} weight_e \tag{7.3}$$

The final result itself is a vector which approaches the calculated fusion point with a predefined speed parameter. If no events remain active in the vector space, the fusion vector approaches a neutral state. The fusion vector serves as an additional means of smoothing: If we can assume that the thought affective state is unlikely to undergo quick changes, the vector can be defined to move slowly, making the algorithm more robust to occasional misclassifications. Also note that the calculation of the fusion point leads to axial dependencies in a multidimensional event space. This may be of interest in a scenario where the positive recognition of a class is meant to have a decreasing effect on the likelihood of another one. If axial independence is needed, the use of several one-dimensional models with subsequent combination is advised.

7.3.2 Gravity Fusion

The term gravity fusion describes a refinement of the previously described vector fusion algorithm. As in vector fusion, recognized events are translated into a vector representation in a (multi-) dimensional vector space, but instead of relying mainly on a decreasing

vector length, gravity fusion interprets single events as mass points with a fixed position. The event vectors, calculated as in vector fusion, hereby define the exact position of these mass points within the vector space. The temporal dynamic of the fusion model is introduced by a temporal decay of the weight of mass points (Figure 7.9). Initial mass of the mass points are defined per event type, position is determined by classification confidence. Based on the current weights of all active mass points, a mass centre can be calculated. The fusion result migrates into the direction of the found centre of mass.

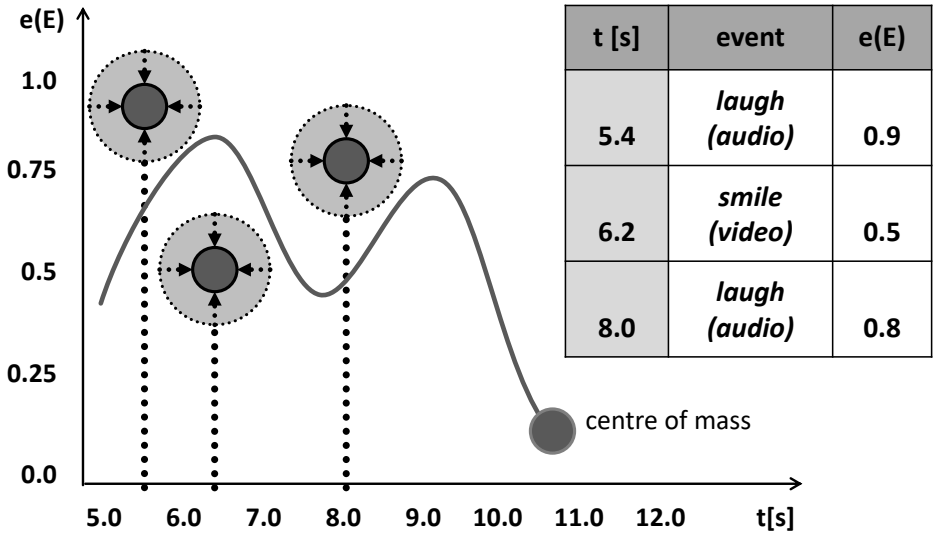


Figure 7.9: Example schematic of the gravity fusion algorithm. Three enjoyment indicating events from audio and video modality are successively mapped into the vector space and resulting vectors (dotted arrows) describe positions of mass points for each event. Weights of mass points decrease over time (shrinking dotted circles), the fusion result migrates in the direction of the centre of mass, which is recalculated every frame.

The fact that in both presented fusion schemes the fusion result does not instantly assume the value of the fusion point (or mass centre respectively), but instead approaches it in a predefined speed, gives the continuous result of event-driven vector fusion a special characteristic: The fusion result reacts inertially to new events. Some misclassifications that happen during a row of correct interpretations do not directly shift the overall result in a wrong direction. On the other hand, this slow reaction time can have negative effects, for example if quick classification switches between classes is desired. A possible countermeasure is to raise the speed of the fusion vector towards the mass centre or lower the lifespan of active events - of course this goes along with lowering the mentioned robustness to single misinterpretations. As a consequence, the decay speed and weights of events have to be adapted to the observed classification problem.

7.3.3 Dynamic Bayesian Networks

While the vector and gravity fusion strategies are custom algorithms designed primarily to deal with the concept of affective events, we will furthermore include a more traditional classification algorithm that features a convenient all characteristics to serve as a means for event driven fusion: Dynamic Bayesian networks (Chapter 3.2.3) are graphical models meant to collect and relate observations from various sources whilst monitoring their temporal flow over prior timeslices. These characteristics perfectly match the requirements of an algorithm that can realize an event-driven fusion strategy.

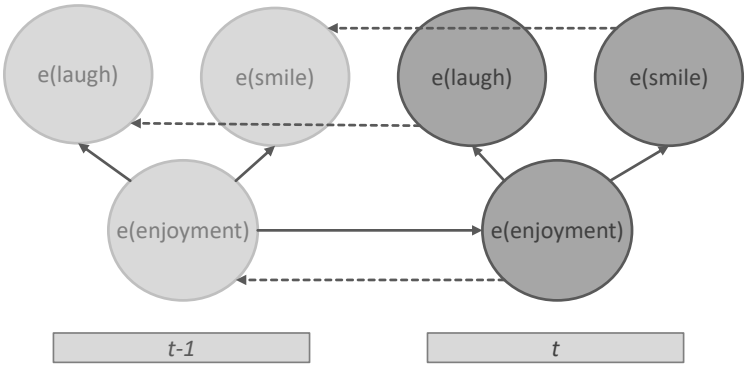


Figure 7.10: Structure of a dynamic Bayesian network for event-driven enjoyment recognition. Each frame event nodes *t* are updated with current events and outdated confidence values are shifted one timeslice into the past *t-1* (dotted arrows). Target class estimation is therefore calculated from current observations and probability distributions of past frames.

Figure 7.10 shows the structure of a dynamic Bayesian network (Chapter 3.2.3), that is used to collect affective events, tagged with respective confidence values. From these it calculates the target class probability based on current and preceding observations: Every frame, modalities are checked for occurring target class indicating affective events. In case of positive recognition, confidence values of event recognizers are used to update the related node within the network. Before this update step, each probability within the present nodes (*t*) is shifted one time slice into the past (*t-1*). Probabilities within the current frame are subsequently calculated from current observations and probability distributions of past frames. Initial configuration is learnt from framewise annotations of the used corpus and models general distributions of frames containing affective episodes.

7.3.4 Possible Advantages of Event-driven Fusion Strategies

The presented approaches to asynchronous and event-driven fusion offer some possible advantages over conventional model-based fusion such as asynchronous hidden Markov models or recurrent neural networks.

- **Reduced Complexity and Enhanced Transparency**
  The main goal to achieve for all model-based fusion approaches is to respect the asynchronicity between affective cues in different modalities. While typical asynchronous fusion strategies theoretically outperform synchronous approaches for this reason, they also have some drawbacks. One disadvantage stems from the inherent complexity of asynchronous models. The parameters a Markov model or neural network has learned and applies to reach a decision are hard to trace, and using them in in real-time systems comes with the risk that trained parameters may not translate well into an uncontrolled environment. Furthermore, once they are trained they can be seen as a black box whose parameters cannot be adjusted to changing conditions. A possible way to make the fusion process more transparent is by shifting from a frame-by-frame based processing towards an event-driven approach. The promise of an event based approach lies in the introduction of an immediate layer of representation that reduces the complexity of the emotion recognition task while enhancing its transparency. An additional benefit of an indirect fusion approach is that these concretely defined interpretation events such as laughs are in some cases easier to detect and classify than a rather abstract emotional category like enjoyment (Lingenfelser *et al.* [111]).

- **Modular Expandability**
  The introduction of events as an intermediate abstraction layer decouples the unimodal processing from the final decision making process. Hereby, each modality represents as a client which can individually decide when relevant information should be provided to the fusion process. Because of this abstraction layer, event recognition components can be added or removed without having to touch the actual fusion system (apart from eventual reconfiguration of some parameters). Missing input from one modality does not cause the the whole fusion process to collapse (Chapter 9). This modular characteristic enables easy expandability - further inclusion of recognition components that supply events shaped to indicate the sought target class is often possible without changing fusion parameters.

- **Multi-user Models**
  The event-driven fusion approach also offers a very convenient concept to handle multi-user scenarios. Sticking to the practical use case of enjoyment detection (Section 7.2.2), one can not only assess the enjoyment level of a single user but can also make statements about the enjoyability of e.g. a conversation between multiple users.

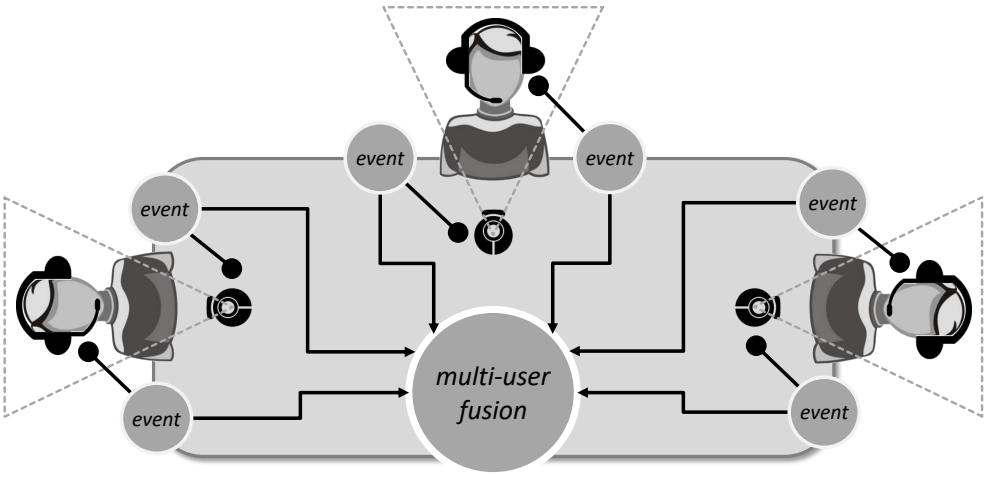


Figure 7.11: Event-driven fusion strategies offer a convenient approach to design multi-user fusion. By processing multi-modal events from multiple participants the state of an interaction can be classified.

The fusion models described in the following section are of course able to process affective events coming from differing persons. By collecting these multi-user events and monitoring their frequency over time we can classify the state of an interaction in addition to affective states of contributors (Figure 7.11).

Figure 7.12: Depiction of frames containing laughs (coloured in black) of multiple users over a one hour recording within a conversational setting. Laughs tend to occur within the whole group and do not often happen in isolation.

The concept to link the affective events of multiple users is reasonable, as in settings such as conversational scenarios participants can influence each other's behaviour. To give an example, Figure 7.12 shows audio frames including voiced

laughs from three users over the course of an hour of conversation. We can see that laughs seldom happen in isolation, but are in this case more of a group phenomenon. Consequently we can derive the enjoyableness of a conversation from these overlapping laughter events and could even go as far as influencing the recognition probability of a single-user laugh based on the observations if other users are currently laughing.

Chapter 8

A Case Example: Enjoyment Recognition with Multi-modal Neural Networks and Event-driven Fusion

As indicated in Chapter 7.2.2, we now choose the recognition of a speaker's level of enjoyment in a natural conversation as scenario for evaluating asynchronous and event-driven fusion approaches. Enjoyment is an important affective state to observe in HCI. Signs of enjoyment, such as laughs and smiles, play a significant role in human communication. Systems that take the role of e.g. companions or tutors should be able to recognize and estimate their presence (or absence) in real-time (or at least with an acceptable delay) in order to design an engaging and entertaining interaction (Niewiadomski *et al.* [127]).

The aim of the presented study (Lingenfelser *et al.* [112]) is to investigate the usability of events as an additional abstraction layer (Chapter 7.2.1), and enjoyment is a well suited recognition task for this investigation: Enjoyment can be defined as an episode of enjoyable emotion. These episodes are typically accompanied by visual and auditory cues that are not by default aligned across channels. Enjoyment episodes can be directly classified from facial expressions or prosodic characteristics extracted from the audio channel. When choosing an indirect approach, we need to apply the intermediate abstraction level of events. Mentioned audio features can be used to classify laugh events and their repeated occurrence can be taken as a hint of enjoyment. At the same time we are able to detect smiles in the face and therefore reinforce the chance of an ongoing

enjoyment episode. This level of abstraction can be applied to any desired modality and found events can consequently be fed into combination algorithms that relate these indicator events back to the target class while considering their temporal flow.

8.1 Belfast Story Telling Database

For the comparison study of asynchronous and event-driven affect recognition strategies we use the first session of the Belfast Storytelling Database (McKeown *et al.* [117]). It features naturalistic and non-acted conversational data between multiple persons. Topics of the conversations are short stories about personal experiences that induced enjoyable emotions within the probands. The described positive emotion of enjoyment is defined to be indicated by visual and auditory cues of enjoyment, such as smiles and voiced laughters.

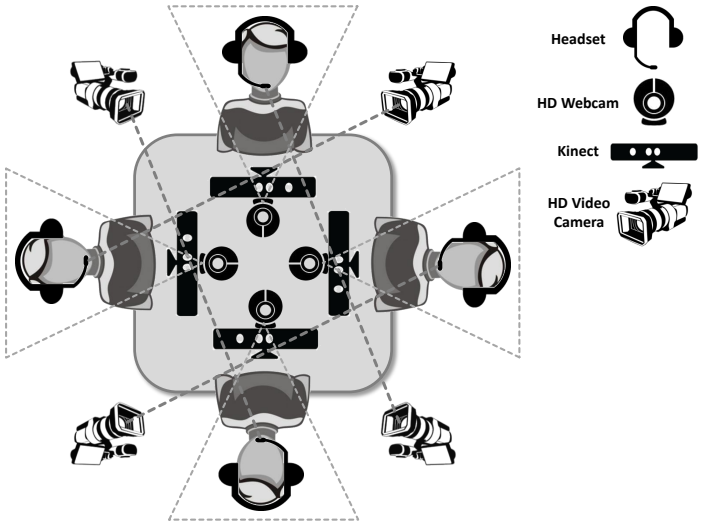


Figure 8.1: Round-table collocation of participants during storytelling sessions, including positioning of HD webcams, Microsoft Kinect^{TM}devices, HD video cameras and head mounted microphones.

The corpus is composed of sessions featuring groups of three to four people sharing stories in either English or Spanish language. Each recording session took about 120 minutes, which results in about 75 minutes of recorded conversation. The applied storytelling task, meant to induce a fluid conversation, is based on the 16 Enjoyable Emotions Induction Task (Hofmann *et al.* [85]). Participants were recruited at least a week ahead of the recording session, and were instructed to prepare or think of stories that relate to each of 16 listed positive emotions or sensory experiences. During the sessions participants were seated in chairs around a central table and wore head-mounted

microphones to capture the verbal information. Video was recorded using fixed HD webcams. Kinect™motion capture technology was used to capture facial features, gaze direction and depth information (Figure 8.1). Synchronisation was achieved using the Social Signal Interpretation framework (SSI) (Wagner *et al.* [169]). The participants took turns at telling personal stories which they associated with an enjoyable emotion. The list of enjoyable emotions (offered by the protocol) was randomised for each story telling session, each round had every participant recall a story connected to the same emotion. The amount of enjoyment naturally varied from story to story. After the story-telling parts, participants occasionally had an unsupervised discussion, which resulted in further natural enjoyment episodes.

Annotations within the Belfast Storytelling Database segment the audiovisual data on a number of different levels (Chapter 7.7). Each story-telling session is segmented to distinguish between story-teller and listeners. There are then laughter segmentations at the two levels visual enjoyment (smile) and auditory enjoyment (laugh). Smile annotations are primarily based on the onset and offset of FACS Action Unit 12 during a laugh episode. Laugh annotation labels the acoustic components associated with laughter; from the onset to offset of audible laugh related sounds during a laugh episode. Persistent accumulations of these enjoyment indicating cues are annotated as enjoyment episodes.

8.2 Topology of Investigated Enjoyment Recognition Systems

Given the naturalistic, multi-person Belfast Storytelling Database and annotations for smiles, laughs and enjoyment episodes (Section 8.1), we can carry out a practical comparison study for the classification methodologies and fusion systems that have been discussed so far. The first step in in every enjoyment recognition strategy is a framewise activity check for modalities with a frame size of 400 milliseconds and 600 milliseconds delta size, resulting in a calculation window of one second and a decision rate of 2.5 Hz. For the audio modality, signal to noise ratio is calculated each frame. We look for coherent signal parts in which the mean of squared input values, multiplied by a Hamming window, exceed predefined thresholds for intensity and length. Meaningful activity in the video modality is assured by the tracking feedback given by the Kinect™device. We check the 100 tracked facial points per frame for valid values and if at least 50% of the 25 frames within the one-second calculation window implicate complete tracking, the feedback is positive for this frame. In order to simulate a working real-time system

evaluation is carried out on full sessions, i.e. end-to-end recognition without exclusion of frames. Consequently, in case of synchronous fusion a decision needs to be forced for frames where no signal is detected (i.e. no face tracked or silence in the audio channel). These frames are mapped onto the class with the highest a priori probability within the training data (i.e. no-Enjoyment).

Each affect recognition system in the presented evaluation uses the same set of features for the audio and video modalities: For vocal analysis we compute acoustic features related to the paralinguistic message of speech, i.e. the features describe *how* something is said, no information about content is included. The feature set used to characterize the audio streams is EmoVoice (Vogt *et al.* [166]). Classifiers for the facial modality are trained with 36 features gained from statistical values over the action unit assessments provided by the Microsoft Kinect™device. Input and target features are standardized to zero mean and unit variance on the training set.

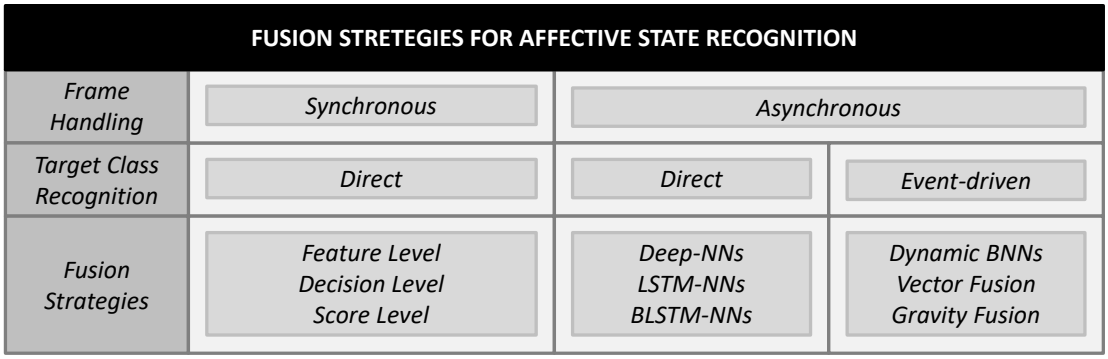


Figure 8.2: We can hierarchically group the investigated fusion strategies depending on the decisions made for the treatment of the temporal dynamics and the levels of processing. The first layer depicts the decision between a synchronous or asynchronous approach. Second, classifiers can be trained for recognizing the intended affective state directly or for recognizing intermediate events in terms of affective cues that are algorithmically interpreted for target class estimation.

In the following sections we will first have a look at the uni-modal baseline systems achieved by applying framewise synchronous and asynchronous enjoyment classification. To lay a foundation for investigating event-driven fusion systems we will evaluate the accuracy of detecting uni-modal laugh and smile events and compare it to direct enjoyment recognition. We will furthermore feed these events into uni-modal event fusion systems as first estimation of the potential gains of the event-driven approach. Afterwards we will apply multi-modal fusion algorithms: Figure 8.2 shows a classification of implemented fusion strategies depending on the treatment of timing and the levels of processing.

8.3 Evaluation of Uni-modal Baseline Systems and Multi-modal Fusion Approaches

We have by now discussed the potential benefits of asynchronous over synchronous fusion approaches and argued in favour of an intermediate layer between low-level features and high-level affective states. In this section, concrete implementations of respective algorithms will be evaluated and compared to uni-modal baseline systems in order to investigate the assumptions made in detail. Result tables report unweighted recognition results (average accuracy across classes), as classified frames contain less samples of occurring enjoyment as well as audible and visible laughs and smiles. Evaluation is user-independent, recordings of a single user are held back as test set, while remaining samples are used for training the respective recognition systems, which leads to a rough total of 18.000 samples for training and 9.000 samples for testing.

8.3.1 Uni-modal Baseline Systems

Table 8.1 shows recognition results for unimodal and synchronous classification - single channel classification with models trained directly on enjoyment annotations. We use a standard support vector machine (SVM) implementation (Chang and Lin [29]) for direct synchronous classification of enjoyment episodes. Recognition of enjoyment via the audio modality is close to random (55.31%). Expressive audible cues for enjoyment are located within the boundaries of an amused episode, but do not fit them very well, which leads to noisy features and poor classification rates (Figure 7.7) of single frames. With an unweighted 71.74%, the video modality yields far better capabilities of determining enjoyment frames. Facial expressions, which express enjoyable emotions, correspond much better to the overarching annotation, as hints of smiles are mostly present during enjoyment.

Another possible approach - without applying a multi-modal fusion strategy - is to consider the temporal flow of observed frames. We use bidirectional long short-term memory neural networks (BLSTM-NN) (Chapter 3.2.4) to model a memory of surrounding frames and therefore incorporate temporal alignments into the classification. It is obvious that direct classification of enjoyment episodes is a demanding task, especially if only using the audio modality. But if we take the ability to use past frames for decision making into account, we are able to greatly increase classification accuracies (Table 8.1). Especially on the audio modality an impressive improvement of 15.21% can be observed. The

Uni-modal Enjoyment Recognition				
	Synchronous		**Asynchronous**	
	Audio	**Video**	**Audio**	**Video**
Enjoyment	50.16%	67.18%	65.41%	76.31%
¬ **Enjoyment**	60.45%	76.29%	75.62%	74.89%
UA	55.31%	71.74%	70.52%	75.60%

Table 8.1: Uni-modal synchronous and asynchronous enjoyment recognition. With synchronous recognition, the video modality corresponds better to the progression of enjoyment episodes than the audio channel. When switching to asynchronous recognition, the recognition accuracy is greatly increased by the consideration of the temporal flow of observed frames.

problem of audible cues not fitting well the boundaries of enjoyment episodes is reduced by asynchronous enjoyment recognition.

8.3.2 Quality of Enjoyment Indicating Events

Next we need to get some insights into the capability of event recognizers for the audio and video modality. Table 8.2 shows the recognition accuracy for laugh and smile events, that can be used to algorithmically derive long-term enjoyment episodes on event level. These are not trained with the bi-modal annotations of enjoyment, but with more narrow uni-modal annotations for actual laughter occurrences and smiles respectively. We use SVM classification for the task of event recognition, as theses short termed cues should be identifiable within a single frame.

Indicator Event Recognition			
	Audio	**Video**	
Laugh	76.51%	78.20%	**Smile**
¬ **Laugh**	91.60%	79.75%	¬ **Smile**
UA	84.05%	78.98%	**UA**

Table 8.2: Indicator event recognition. Short-termed events are easier to classify than the abstract affective target class.

84.05% accuracy for audible laughs and 78.98% for visual smiles respectively give an impression of the easing on classification difficulty if recognizers are trained on

the recognition of short-term events. The high classification accuracy of laugh frames is of special interest, as the gap between the recognition of affective hints and direct affect classification is particularly high in this case (84.05% to 55.31% and 70.52% respectively). Consequently, fusion approaches that are designed to make use of event recognition should be able utilize audible information to the fullest.

8.3.3 Uni-modal and Event-driven Enjoyment Recognition

As a last uni-modal experiment, we are able to apply event-driven recognition approaches to one single modality, by relating recognized events of a single channel back to whole affective episodes via event-driven fusion schemes (Figure 8.3). Most likely, better results can be expected when events of multiple modalities are fed into the fusion process. But the appliance of the event-driven approach already shows encouraging results in a uni-modal scenario (Table 8.3).

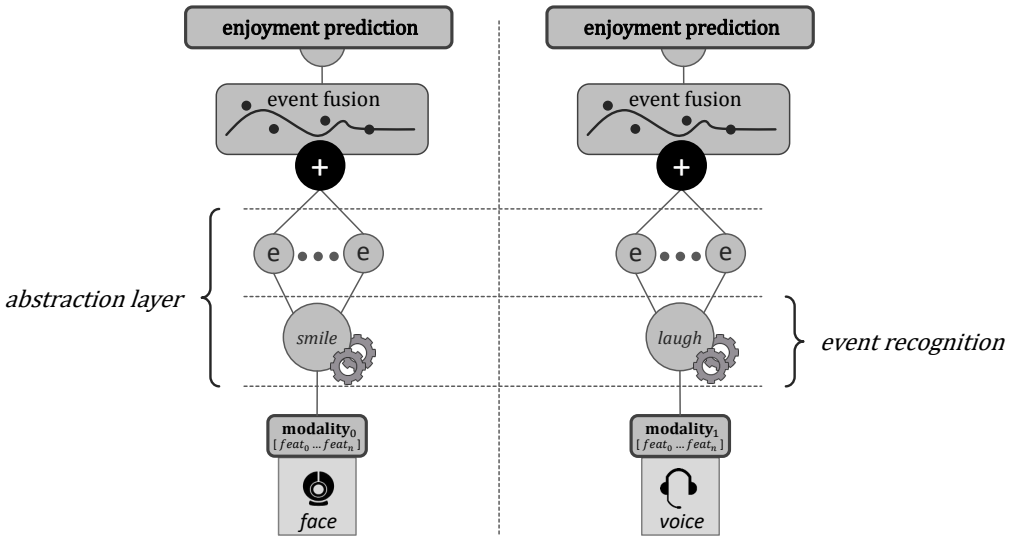


Figure 8.3: Event-driven fusion approaches can be used combine to indicator events of a single modality. Event recognition and abstraction layer are only applied to one signal source. We are eventually not able to catch whole enjoyment episodes and better results are possible when events of multiple affective channels are fed into the fusion process - first assertions about the quality of an event-driven approach are however possible.

Deriving sought affective episodes from audible laughs with the vector fusion approach, we are able to achieve an accuracy of 74.70% for classification of enjoyment frames - an improvement of 4.18% compared to direct asynchronous enjoyment recognition. As discussed before, the audio channel can apparently best be employed on the event level.

Uni-modal Event-based Enjoyment Recognition		
	Audio	**Video**
Enjoyment	78.45%	80.13%
¬ Enjoyment	70.95%	71.39%
UA	74.70%	75.76%

Table 8.3: Event fusion algorithm applied to uni-modal events. The indirect approach already results in improved enjoyment classification accuracies, without taking multi-modal information into account.

Taking only smile events from the video channel into account results in a recognition rate of 75.76%, almost identical to the asynchronous uni-modal baseline system.

8.3.4 Synchronous Fusion Systems

All so far discussed affect classification approaches have only made use of direct or event-based information from a single modality source. From this point on, we will analyse results that base on the combined insights gained from multiple channels. We will first start with synchronous fusion schemes, the most simple and common approaches to multi-modal fusion. Asynchronous fusion systems apply classification models that are suited to catch temporal alignments between observed modalities better than their synchronous counterparts. The last group of algorithms that will be discussed are the event-driven fusion approaches, that rely on the indirect recognition of target class indicating cues and the modelling of their temporal course during enjoyment episodes.

Feature, decision and score level fusion (Chapter 4) are obvious approaches to combine multi-modal information from different sources as these algorithms can be implemented in a straightforward manner, work on the basis of synchronous combination of channels and use direct classification results. Therefore they are applied in most studies dealing with multi-modal affect recognition and can serve as a baseline for more elaborate recognition schemes. In addition to the feature fusion algorithm, several representative fusion schemes for decision and score level have been tested with very close average recognition rates. Presented results are generated with the product rule (decision level - Chapter 4.1.3) and stacking (model level - Chapter 4.1.4). Synchronous SVM classification is used in order to fully exclude the temporal aspect.

Table 8.4 shows that discrepancies between enjoyment classification on the audio and video modality (Table 8.1) pass on to these simple synchronous fusion approaches: Fea-

Synchronous Fusion			
	Feature	**Decision**	**Score**
Enjoyment	63.14%	57.39%	36.14%
¬ **Enjoyment**	59.08%	74.32%	88.20%
UA	61.11%	65.86%	62.17%

Table 8.4: Synchronous fusion of direct enjoyment classification results of the audio and video modality. Poor results of enjoyment classification of the audio channel fully contribute to the fusion result.

ture, decision and score level fusion perform on an intermediate level between the merged modalities (61.11%, 65.86% and 62.17%). This is to be expected, as the problematic enjoyment classification models trained on the vocal modality fully contribute to the fusion result.

Given these rather bad results (we observe drops in recognition rates below synchronous uni-modal baseline systems), we can at this point introduce an experimental intermediate step between synchronous and event-driven fusion: We can utilize the event classification models trained on uni-modal annotations for enjoyment indicating cues (laughs and smiles) and apply them directly in decision and score level fusion schemes: We replace models trained on enjoyment episodes with classifiers meant to detect audible laughs and visual smiles and map the corresponding probabilities to the enjoyment classes.

Synchronous Fusion with Mapped Events		
	Decision	**Score**
Enjoyment	55.16%	66.75%
¬ **Enjoyment**	90.32%	80.54%
UA	72.74%	73.65%

Table 8.5: Synchronous fusion with mapped events. As an experimental intermediate step we can utilize the event classification models trained on uni-modal annotations for enjoyment indicating cues (laughs and smiles) and apply them directly in decision and score level fusion schemes.

Results of this procedure are described Table 8.5. The experimental synchronous fusion with mapped events delivers good results on decision as well was on score level. With a recognition rate of 72.74% on decision and 73.65% on model level they exceed the synchronous uni-modal baseline systems, but stay behind asynchronous enjoyment classification on the video channel. By combining mapped laughter and smile detections,

these approaches are able to partially capture the course of enjoyment episodes, but they do not take temporal relations of recognized events into account. We see that the main improvements in recognition performance is based on the detection of ¬Enjoyment. This means they mostly predict the absence of indicating events during the periods of enjoyment, there are still many misclassifications.

8.3.5 Multi-modal Neural Networks

For asynchronous fusion we use a direct feature fusion approach to affect recognition (Figure 8.4). But instead of synchronously considering the multiple channels, the inherent logic of recurrent neural networks (Chapter 3.2.4) should be able to catch asynchronous relationships between modalities. By including multi-modal information into the asynchronous classification schemes, the characteristics of an enjoyment episode can be adequately modelled.

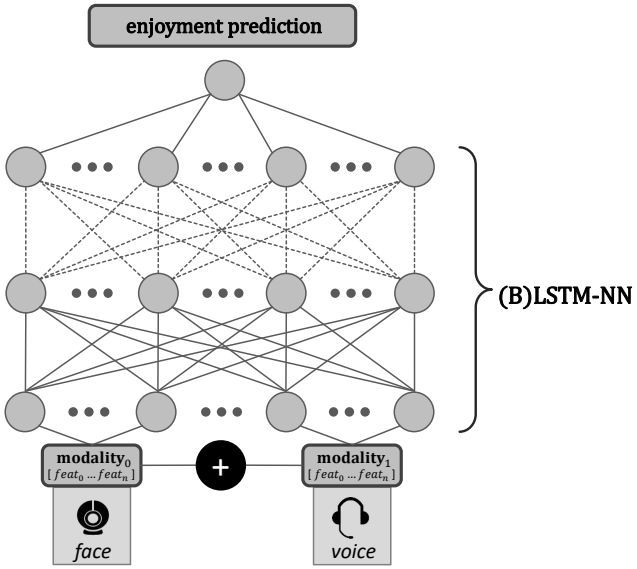


Figure 8.4: The multi-modal neural network merges affective channels on the feature level. Hereby, the memory capabilities of recurrent neural networks (LSTM-NN and BLSTM-NN) should enable the system to adapt to temporal asynchronicities.

Especially the long-short-term memory cells of LSTM and BLSTM neural networks seem to be able to capture the temporal dependencies between affective cues within the observed modalities and result in enjoyment classification rates of up to 75.76%, a very minor improvement compared to asynchronous enjoyment recognition with the video modality only.

Direct Asynchronous Fusion			
	Deep-NN	**LSTM-NN**	**BLSTM-NN**
Enjoyment	82.29%	61.99%	67.15%
¬ Enjoyment	51.27%	86.16%	84.37%
UA	66.78%	74.08%	75.76%

Table 8.6: Asynchronous fusion. The memory capabilities of recurrent algorithms enable the capture of temporal dependencies between observed channels.

All neural networks were trained using the stochastic gradient descent (SGD) algorithm with a momentum value of 0.9. We used grid search to determine the optimal learning rate (1e-1, 1e-2, 1e-3, 1e-4), number of hidden layers (1, 2, 3), and number of hidden nodes (128, 256, 512, 1024). Each hidden layer has the same number of hidden nodes. In order to find the needed parameters, we held out roughly half of the training samples training samples for validation. Hyper-parameters were tuned on the validation set until the validation error did not improve for at least 10 epochs and we choose the networks that achieved the best validation error. The validation set was then combined with the training set and we retrain the network on the combined data. For audio-related features, principle component analysis (PCA) is further used to reduce the dimensionality and therfore greatly reduce the high computational cost. For PCA, we retain 95% of the variance. Besides, when training BLSTM-NNs, we added Gaussian noise with zero mean and standard deviation 0.1 to the inputs. Besides, sequences and fractions were shuffled randomly. In our experiments, we trained DNNs, LSTM-NNs and BLSTM-NNs using the open-source implementations Theano (Bergstra *et al.* [13]) and CURRENNT (Weninger *et al.* [175]).

8.3.6 Event-driven Fusion Systems

We have by now seen that classification models for short termed enjoyment indicating events (in our case) perform more reliable than the direct recognition of enjoyment (Table 8.1 versus Table 8.2). Furthermore, we can state that an asynchronous treatment of signal frames that considers (in the case of multi-modal fusion) temporal relations between modalities outperforms synchronous approaches (Tables 8.4 to 8.6). In this final evaluation we will now try to combine these identified benefits with event-driven fusion approaches (Figure 8.5).

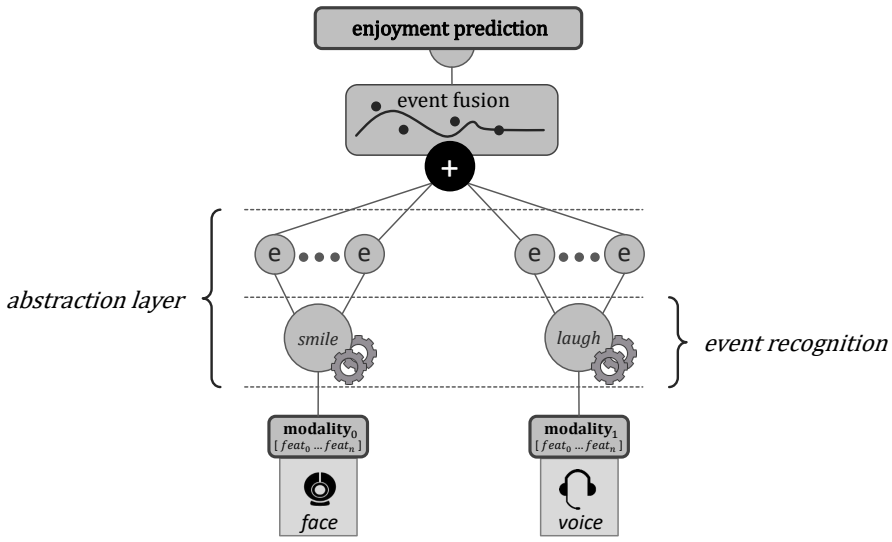


Figure 8.5: In the final event-driven fusion system we do not classify enjoyment episodes directly. Instead we combine the enjoyment indicating laughs from facial and vocal modalities (smiles and laughs).

Event-driven Fusion			
	Dynamic-BN	**Vector**	**Gravity**
Enjoyment	78.45%	70.50%	78.52%
¬ Enjoyment	71.77%	84.91%	80.51%
UA	75.21%	77.70%	79.51%

Table 8.7: Event-driven fusion. By combining the recognition of enjoyment indicating short term events and the possibility to temporally relate these multi-modal events, event-driven fusion schemes achieve the best performance of the comparison study (in this case shown by the gravity fusion model).

Bringing together the recognition of enjoyment indicating short term events and the possibility to temporally relate these multi-modal events, event-driven fusion schemes achieve good recognition rates with the best performance in this case shown by the gravity fusion model with 79.51%. This is the best result we were able to achieve for enjoyment recognition during our experiments with the examined approaches. According to McNemar's Chi-Squared Test ($p < 0.05$), improvements in comparison to the second best approach (vector fusion with 77.70%) are significant. Table 8.7 also shows a well balanced distribution of accuracies among classes (78.52% for Enjoyment and 80.51% for ¬Enjoyment). The results demonstrate that event-driven fusion is accurate in classifying whole episodes of enjoyment and models their boundaries well.

Initial vector lengths (vector fusion) and mass points (gravity fusion) respectively are

directly derived from probabilities given by the event recognizers. This derivation only makes sense if confidence values of given classifiers are comparable. To prove this assumption, Figure 8.6 plots the confidence values of event recognizers against the correctness of the estimation. Prediction behaviours of modalities resemble each other clearly.

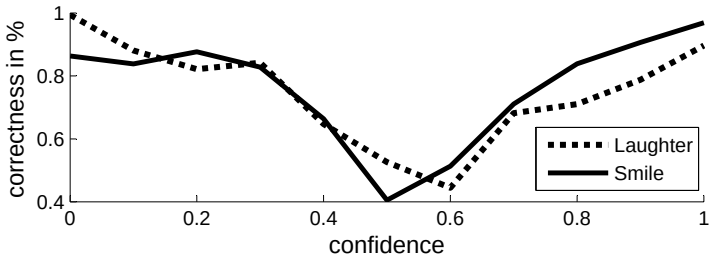


Figure 8.6: Frequency of correctly classified frames according to laugh / smile confidence. Similar prediction behaviour allows to directly combine confidence values during the fusion process.

Optimal configuration of parameters have been empirically determined by systematically testing a large number of combinations (Figures 8.7) with a grid search approach during the training phase.

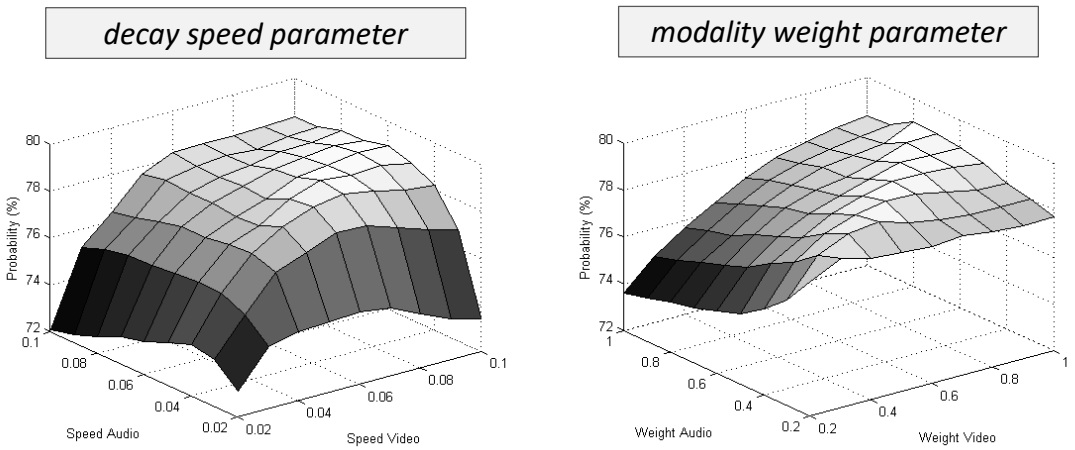


Figure 8.7: Influence of audio and video event decay speed and modality weights on vector fusion performance. Stable performance is observed if smiles have a high decay speed compared to laughs and audio and video events are weighted in a ratio of 8 to 10.

The most important parameter for vector and gravity fusion is the decay speed of registered events i.e. how long an event will sustain influence on the fusion result. Speed of smile events is regulated high as the beginning and ending of enjoyment is often characterized by the presence and absence of smiles in the face. Consequently, the fusion

vector should rise and fall fast whenever smiles are recognized or not. The decay speed of laughter events is regulated low. Laughs are considered a strong indicator of enjoyment and whenever they occur we expect the enjoyment episode to last for several frames afterwards. Best performance for the gravity fusion algorithm is achieved if laugh events are by default weighted less than smile events - again due to the fact that smiles better describe the limits of enjoyment segments.

8.4 Evaluation Summary and Conclusions

In Table 8.8 we give a summary of the results presented in the comparison study. Recognition accuracy of multi-modal fusion approaches are compared to uni-modal baseline systems and successive improvements (or degradation respectively) are calculated in the *effect* column.

Summary of Results			
Approach	**Algorithm**	**Result**	**Effect**
	Unimodal		
Synchronous	Video SVM	71.74%	**-**
Asynchronous	Video BLSTM	75.60%	**+ 3.86%**
Event-driven	Video Vector Fusion	75.76%	**+ 4.02%**
	Multimodal		
Synchronous	Decision Level	65.86%	**- 5.88%**
Asynchronous	BLSTM-NNs	75.76%	**+ 4.02%**
Event-based	Gravity Fusion	79.51%	**+ 7.83%**

Table 8.8: Comparison of tested approaches to affect recognition, differentiated in relation to the classification of models in Figure 8.2. Results of respective best performing algorithms are shown with the achieved effect in relation to uni-modal, synchronous enjoyment classification on the video channel.

The first baseline approach to enjoyment recognition is synchronous uni-modal classification achieved with features from the video modality - as direct enjoyment classification on the audio channel yields very low accuracy (Table 8.1). While enjoyment classification on the basis of Kinect^{TM}facial features showed an acceptable recognition accuracy, the audio modality stayed on a close to random level. By applying asynchronous classification techniques via neural networks with memory capabilities, these results can

be raised significantly in both modalities. An improvement of 3.86% can be achieved if we switch to asynchronous classification of enjoyment on the video channel (Table 8.1). The consideration of temporal alignments together with good recognition rates for event recognition (Table 8.2) leads to an even more accurate classification rate for the uni-modal event-driven approach (Table 8.3). By combining smile events found in the video modality over time with the vector fusion algorithm, the uni-modal result can be raised by 4.02%. Note that the term *fusion* does not completely fit in this case as we use it mainly for the combination of several modalities.

First impressions gained for uni-modal classification on the better suited video channel carry over to multi-modal fusion schemes. When synchronously taking information from multiple channels into account, the common strategies of feature, decision and score level fusion are tested as a first step. On average, these algorithms result, however, into a performance that lies between the classification rates of the single channels (Table 8.4). Because of the bad performance of the audio modality, the averaged accuracy lies 5.88% below the uni-modal baseline using only the video channel. The shortcomings of synchronously classifying audio frames directly influences the performance of the widely used synchronous fusion approaches and results in a mediocre recognition accuracy. The neural network based asynchronous fusion systems stick to direct classification, but introduce memory cells to model the asynchronous dependencies between modalities (Table 8.6). These dedicated approaches are able to better catch the multi-modal characteristics of enjoyment episodes and therefore yield recognition enhancements of up to 4.02%. Event-driven fusion schemes combine an indirect classification approach with algorithms able to consider the temporal relations between the recognized multi-modal events (Table 8.7). This combination leads to an improvement of recognition accuracy for enjoyment frames of 7.83% with the gravity fusion algorithm compared to the uni-modal base system and a 13.71% higher accuracy than the badly performing synchronous fusion strategies.

Taking this discussion in a more general direction, we can state that affect recognition systems apply multi-modal fusion under the reasonable assumption that combination of information from several modalities does improve classification accuracy. However, studies presented in preceding chapters showed that the real enhancements of fusion systems compared to uni-modal classification are - to say the least - unstable. If we look at the common synchronous fusion approaches, we in fact observe a severe drop in accuracy compared to the best synchronous single channel enjoyment classification. If the analysis had stopped at this point, one could conclude the failure of multi-modal fusion in this case. However, there are several options to enhance the processing of available information. On

the one hand, we have the option to incorporate information on the temporal alignment into the classification process. Whether we apply asynchronous recognition at the uni-modal or at the fusion level, we observe significant improvements over synchronous approaches. On the other hand, there is an option to lift the classification task on a higher abstraction level. Instead of classifying enjoyment directly, we look for events of smile and laughs and relate these short term indicators back to whole enjoyment episodes. The used algorithms incorporate the temporal dynamics of events and can therefore by classified into the group of asynchronous and model-based approaches.

Inclusion of temporal observations as well indirect classification via event recognition have proven to enhance the performance of examined enjoyment recognition systems and therefore both techniques can be advised for practical applications (which will be reported on in Chapter 10). Event-based fusion strategies applying these techniques also fulfil the requirements demanded by latest considerations about innovative fusion systems (Glodek *et al.* [74]): They are able to compensate for temporarily unavailable data, use information of temporal alignments and are easy to extend to further modalities and event types.

Chapter 9

Multi-modal Fusion in the Wild

Aside from developing fusion methods that are able to process the multi-modal affective cues emitted by humans in a realistic manner, there is the goal to bring these recognition systems to application in real-world scenarios. The efforts made to realize this goal can be subsumed under the lately emerged term *Emotion Recognition in the Wild*. Over the last years, there is common agreement, that affect recognition systems developed in laboratory settings do not translate well into real-life applications because of hindrances experienced in uncontrolled settings such as background noises, bad lighting conditions or the general unavailability of certain sensor devices in a non-stationary, mobile setting. Aware of these problems, databases with real world (or at least more close to) settings (e.g. Static Facial Expressions in the Wild (Dhall *et al.* [40]) or Acted Facial Expressions in the Wild (Dhall *et al.* [41])) start to tackle the problem as well as a series of grand challenges (ICMI: Emotion Recognition in the Wild (Dhall *et al.* [42–45]), EmotioNet Challenge (Benitez-Quiroz *et al.* [12])).

In realistic scenarios there are great opportunities for multi-modal fusion systems. However, studies that have focused on the fusion of multiple channels often start from too optimistic assumptions, i.e. that all data from the observed modalities is non-corrupted and available for processing at any given point in time (Figure 9.1). Whether this problem is caused by technical problems like tracking issues or by asynchronous emotional expressions throughout affective channels as indicated in by the segmentation problem (Chapter 6.4), it needs to be addressed by affect recognition systems. Especially event-driven fusion strategies (Chapter 7) are not dependent on constant input from all modalities but work on events that are currently available. As a consequence their functionality is not substantially threatened if a modality remains silent at times. We will elaborate on this assumption in Section 9.2. As their nomenclature suggests, synchronous fusion

approaches typically expect synchronously available input at each decision step. If these inputs are not fully given, the algorithms are prone to pointless decisions or even malfunctions. Therefore ways to improve these algorithms to better deal with these issues need to be found.

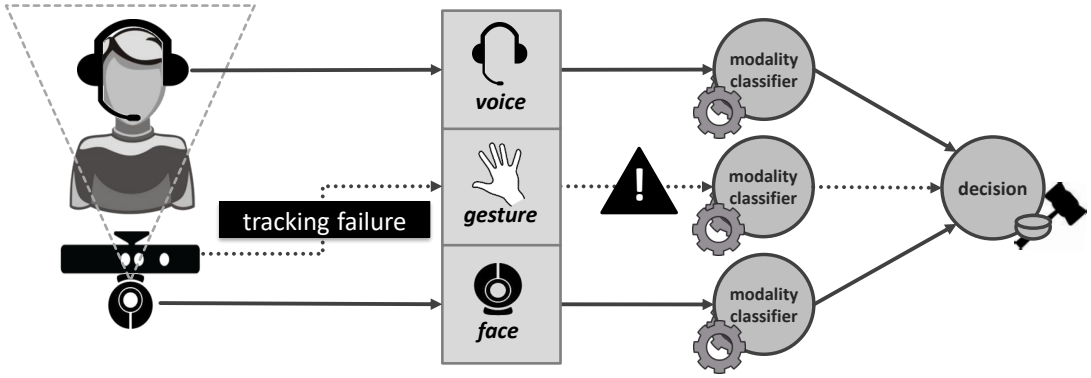


Figure 9.1: Temporal failures in modalities due to problems such as hardware breakdowns, operating errors or tracking problems are to be expected in practical applications.

9.1 Gearing Synchronous Fusion Systems towards Application in the Wild

There are various reasons for missing data to occur. Technical problems such as failing sensor devices can deny the availabality of the respective signal stream. Given a correctly operating sensor we face the possibility of unaccessible information, e.g. a tracked object disappearing from a fixed camera view (Figure 9.1). Even if desired information is at hand, it may be corrupted to a certain degree, e.g. variuous noise sources overlaying a speech signal. In Chapter 6.4 we have already suggested that non-technical problems can be held responsible for one or more modalities to become non-usable: If a user simply does not show meaningful expressions, no relevant statements can be made. We can for example think of a motionless user which will eventually result in no relevant contributions from the gesture modality. Emotional reactions to affective stimuli may occur time shifted across expressive channels. It is therefore of high importance for a robust recognition system to be able to decide dynamically which channels are able to provide usable information. In the follwing evaluation we will first focus on the correct treatment of temporarily missing missing data sources, which corresponds more to technical reasons and is meant to guarantee the robustness of multi-modal recognition

systems. Strategies to tackle the asynchronous emergence of affective cues during emotional episodes has been exhaustively discussed in detail in Chapter 7. Given the case of partially missing data streams, multiple solutions have been proposed: Pantic and Rothkrantz [131] suggest prediction of missing values on basis of previous samples. When correctly detected, unreliable features can be marginalized in order to reduce their influence on the final decision (Demiroglu *et al.* [39]).

As long as a synchronous fusion system is only evaluated on offline data the assumption of complete data in all modalities can be ensured by examining given samples beforehand and excluding parts where one or more channels are corrupted or not usable for evaluation. In application oriented online systems however, we cannot neglect the problem of missing data and need to implement adequate solutions to handle it in order to guarantee recognition robustness. Some studies like (Sebe *et al.* [150]) highlight the benefits of fusion mechanisms in situations with noisy features or missing values of features. Nevertheless, surprisingly few fusion approaches explicitly address the problem of non-available information. Most of them are based on the assumption that all data is available at all time. As a result we conducted the following study (Wagner *et al.* [168]) with the goal to enhance various synchronous fusion techniques with the capability of dealing correctly with temporarily unavailable modalities. The presented systems try to solve the missing data problem at the multi-modal fusion level and implement ensemble strategies with solutions to compensate temporarily unavailable modalities.

9.1.1 Handling Missing Data in Synchronous Fusion

Concerned with systems for multi sensor data fusion and real-time applications, these can be implemented in a way that they resist the breakdown of one or more attached sensors (Figure 9.2). If the classifiers involved in decision making each represent the observations of an associated sensory device, the absence of a single contribution to the final decision is unlikely to result in a drastic quality fall-off for overall classification accuracy - especially if the sensory malfunction is recognized and the corresponding classifier's (most likely counter-productive) contribution is accordingly rated.

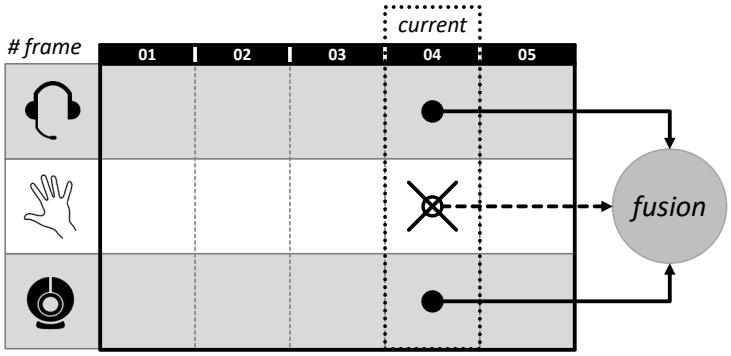


Figure 9.2: The fusion step is a very convenient level within an affect recognition system to handle temporarily missing information from single modalities. Synchronous fusion approaches are originally designed to expect complete input from all data sources and need to be enriched with strategies on how to handle missing data streams.

To achieve this goal, we need a recognition system that handles temporarily[1] missing modalities in the fusion step. As we for now focus on common synchronous fusion approaches that are intrinsically designed to expect equal clocked and complete input from all data sources, we enrich them with strategies on how to handle missing data streams. For the explanation of reviewed algorithms the following annotations are used: The decision of ensemble member t for class n is denoted as $d_{t,n} \in \{0, 1\}$, with $t = 1..T$ and $n = 1..N$ and $d_{t,n} = 1$ if class ω_n is chosen, $d_{t,n} = 0$ otherwise. Respectively the support given to each class n (i.e. the calculated probability for the observed sample to belong to single classes) by classifier t is described as $s_{t,n} \in [1..0]$.

- **Weighted Majority Voting**
  Majority voting simply sums up decisions of T classifiers. The ensemble decision for an observed sample x is chosen to be the class ω_n which received the most votes (decisions) v_n. A definite decision is only guaranteed if an odd number of ensemble members handle a two-class problem. In weighted majority voting each vote is associated with the pre-calculated weight of the ensemble member. The ensemble decision for an observed sample x is chosen to be the class ω_n which received the most weighted votes v_n. Ties are not likely to happen this way, which makes the weighted variant more suited for practical application.

  $$v_n(x) = \sum_{t=1}^{T} w_t d_{t,n}(x)$$

[1] Temporarily means that we adjust the fusion scheme per input sample (not as in most fusion approaches for the whole corpus), i. e. for each modality and each sample, we decide whether to include the information to the fusion process or not. A the moment this decision is exclusive. If we had some way of estimating the degree of corruption within the sample's modality, we could simply assign proper weights instead.

Handling Missing Data:
Ensemble member t containing training data from modalities not featured in an observed sample is not included in the poll.

- **Weighted Average**
  In contrast to weighted majority voting, the weighted average strategy applies weights not to class labels, but to continuous outputs of ensemble members. By summing up the weighted support given to each class ω_n, total weighted support μ_n for class n is calculated as:

$$\mu_n(x) = \sum_{t=1}^{T} w_t s_{t,n}(x)$$

  The ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.

  Handling Missing Data:
  Ensemble member t containing training data from modalities not featured in an observed sample is weighted with a value of zero.

- **Maximum Rule**
  The strategy simply chooses the maximum support generated by T ensemble members. The ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.

  Handling Missing Data:
  Ensemble member t containing training data from modalities not featured in an observed sample gives support s_t of value zero to each of the n classes.

- **Sum Rule**
  The sum rule simply sums up the support given to each class ω_n in order to generate total support μ_n for each class. The ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.
  In the normalized version, the support given to each class ω_n is averaged ($\frac{1}{T}$ serves as normalization factor) and the total support μ_n for class n is calculated as:

$$\mu_n(x) = \frac{1}{T} \sum_{t=1}^{T} s_{t,n}(x)$$

Handling Missing Data:
Ensemble member t containing training data from modalities not featured in an observed sample gives support s_t of value zero to each of the n classes.

- **Product Rule**
  By multiplying the support given to each class ω_n, total support μ_n for class n is calculated as:

$$\mu_n(x) = \frac{1}{T} \prod_{t=1}^{T} s_{t,n}(x)$$

  Note that this fusion strategy reacts very sensitive to pessimistic ensemble members, as a support of value zero virtually nullifies the chance of a class to become the final decision.

  Handling Missing Data:
  Ensemble member t containing training data from modalities not featured in an observed sample gives support s_t of value 1 to each of the n classes.

- **Cascading Specialists**
  The cascading specialists method (Lingenfelser *et al.* [109]) was introduced as a custom fusion algorithm for gaining ensemble decisions in Chapter 5.1.3. It does not focus on merging outputs from all ensemble members, but on selecting experts for each class and bringing them in a reasonable order to support weakly recognized classes and achieve a flattening effect among class accuracies.

  Handling Missing Data:
  The concept of choosing experts for certain classes has to be broadened, so that an expert unable to handle the given sample (because of missing data) can be adequately replaced. Instead of selecting one single ensemble member for expert and final classification tasks, ordered lists containing all classifiers - ranked by their qualification for the given task - replace sole classifiers. If in the classification step missing data is detected and the most qualified ensemble member is trained with data of that type, we move down in the prepared list to find the next best classifier that is able to handle the observed sample.

- **Emotion-adapted Fusion**
  The generic approach to classification in a multi-class environment is to train classifiers and corresponding ensembles to categorise among the available classes, but the structure the chosen emotion model - consisting of two scales for valence and arousal - more emotion-adapted techniques for finding an ensemble decision can be applied (Kim and Lingenfelser [97], Wagner *et al.* [172]). Therefore we chose to include the most promising of emotion-adapted fusion algorithms developed in Chapter 5.2.3 in this list. In these emotion-adapted fusion approaches, several ensembles are trained to recognise the observed emotion's axial alignment. Resulting outputs are logically combined for final decision, as decisions on valence and arousal orientations explicitly describe one of specified classes.

  Handling Missing Data:
  In our implementation, weighted majority voting is used to generate the ensembles' decisions on axial alignments. Therefore the handling of missing data described for respective fusion scheme is adopted.

9.1.2 Experiments on Recordings with Actually Missing Data

For testing introduced fusion systems against the problem of missing data, we first need a corpus in which data is actually temporarily inaccessible. Most freely available corpora have already undergone a full preprocessing and feature only samples with complete data available. Luckily we are able to use a sub-corpus of the natural CALLAS dataset (Caridakis *et al.* [26]) we introduced in Chapter 6.2 for our purpose: It was originally designed for examination of cultural differences between emotion expressions of persons from different European countries and so it features participants from Greece, Italy and Germany. The evaluation bases on data collected from German participants, as we do not aim at dealing with cultural differences in detail. The German sub-corpus contains 21 persons (ten female and eleven male) and almost five hours of recorded interaction, which is sufficient for our investigations. Together with audio and video recordings, the corpus offers a gesture modality which is partially missing due to tracking problems or the simple absence of hand movements. In addition, the facial modality is missing at points in time, when facial recognition module lost track of a recorded person and therefore no meaningful facial features could be extracted. Handling of missing data is modelled within the multi-modal fusion process. In order to recognize missing data, we keep track of when the facial recognition module looses the bounding box around an observed face - this happening marks recorded data as missing until the face is recognised again. We

furthermore introduce a threshold for minimum energy within a signal recorded from the gesture analysis component and whenever energy falls below this mark, we assume that no gesture was performed at all during the recorded phrase.

As explained in Chapter 6.2, users were asked to utter expressive sentences and accompany them with whatever gesture or voice they felt to be fitting. Though used mood inducing sentences are more categorised along the valence axis (positive, neutral and negative), it becomes obvious that an arousal categorisation is also needed. Especially when looking at negative sentences, there are samples tending to a depressed and sad mood (*negative-low*), while others are expressed in an aroused and angry way (*negative-high*). Nearly all neutral and part of the positive observations share a calm and optimistic sub-tone (*positive-low*) in contrast to fewer positive examples bearing clear hints of joy and laughter (*positive-high*). Based on these impressions we refrained from including neutral moods as an own class and label calm and non-negative emotions as *positive-low*.

Features for audio and video channels are similar to experiments in Chapter 6.3. The features we extract from the acceleration signal of hand movement (Figure 9.1) go back to a set of expressivity parameters originally defined by Hartman *et al.* in order to synthesize expressive gestures to be used by conversational agents (Hartmann *et al.* [81]). Caridakis and collegues have applied similar features to measure gesture expressivity in hand tracking from video images (Caridakis *et al.* [24]). They propose six expressivity parameters (overall activation, spatial extent, temporal, fluidity, power/energy and repetitivity) to describe the level of expressivity. Overall activation, for example describes the amount of hand movement in general, while fluidity is a measure to differentiate between smooth and hasty gestures.

modality	**short-term feature**	**long-term feature**	**total**
gesture	3 axes acceleration and 1st derivatives, velocities, positions	power, fluidity, volume, mean, minimum, maximum, position minimum/maximum, length	75

Table 9.1: Overview of feature extraction methods applied to the gesture modality.

To sum up, our corpus with actually missing data features 2513 samples with missing data in the facial and gestural modality. The vocal modality is always accessible, as samples represent one spoken sentence. Facial features can successfully be extracted throughout 2251 observations (90%) and significant gestures are available for 569 samples (24%). In consequence of the experimental design the samples are equally distributed among the 21 participants, however, if we take a closer look at the number of samples per emotion we

find a high variety between the users. In general, female users recorded in this experiment seem to perform more expressive than their male counterparts.

Single Channel Performance			
	Audio	**Video**	**Gesture**
Positive-Low	61.0%	45.0%	57.0%
Positive-High	50.0%	72.0%	30.0%
Negative-Low	49.0%	31.0%	44.0%
Negative-High	43.0%	43.0%	36.0%
UA	51.0%	48.0%	42.0%

Table 9.2: Single channel performance for each modality. The vocal modality outperforms facial and gestural cues. Positive emotions are recognised better than negative ones.

In Table 9.2 single channel performance is shown for each modality. Whenever a sample contains missing data for respective modality, the sample is of course not included for evaluation, so these results stem from different quantities of samples. The vocal modality outperforms facial and gestural cues and establishes most balanced accuracies among observed classes, though positive emotions are recognised better than negative ones. The facial modality recognizes the *positive-high* emotion very well - presumably because it is well suited for detection of smiles and movements of the face associated with laughter - but lacks on other classes. Gestures are most often correctly classified during *positive-low* phases, the most calmly expressed class of observed emotional states with nearly no movement at all. *negative-low* emotions were often expressed with despaired gestures, that could partially be separated from gestures with high arousal. These expressive movements were obviously often misinterpreted among each other, leading to very low accuracies on classes with highly aroused emotions.

Table 9.3 lists results of generic fusion approaches that can be applied to any given classification problem. Theoretically do all decision level fusion strategies aim at exploiting mentioned differences in single channels in order to enhance combined performance. Practically, performances of facial and gestural modalities are too inferior to the audio channel to result in greater gains in overall recognition rates. If one compares the vocal modality to the fusion schemes, most approaches perform better on the *positive-high* class. This behaviour can be explained by the very good result of the facial modality in this area and the resulting influence on the ensemble. Unfortunately do bad results for remaining classes effect overall performance in a contrary way and these gains are

Generic Decision Level Fusion						
	Voting	**Avg**	**Max**	**Sum**	**Prod**	**CS**
Positive-Low	61.0%	53.0%	55.0%	57.0%	58.0%	49.0%
Positive-High	52.0%	63.0%	65.0%	66.0%	66.0%	63.0%
Negative-Low	49.0%	39.0%	36.0%	41.0%	42.0%	47.0%
Negative-High	42.0%	41.0%	46.0%	37.0%	38.0%	40.0%
UA	51.0%	49.0%	50.0%	50.0%	51.0%	50.0%

Table 9.3: Fusion results on samples with partially missing data with generic decision level fusion approaches. No strategy generates drastically worse results than the best modality - a good indication that introduced strategies to handle missing data lead to robust recognition results.

lost again in other categories. This can be well observed when looking at the product (Prod) and sum rule (Sum) - the ”standard” fusion schemes for merging classifier outputs - as recognition results stabilize around vocal performance with a better trend on the second class. Same estimations hold for other merging strategies such as the weighted average rule (Avg). Behaviour of approaches that choose exactly one support value among ensemble members for each class like the maximum rule (Max) resembles the just mentioned characteristics. Weighted majority voting (Voting) has an inherent weighting method that causes a strong reliance on the dominant modality, resulting in almost the same accuracies across all classes as the audio channel. The cascading specialists (CS) strategy generates acceptable results on negative classes - that are all in all more weakly categorised throughout the ensemble - but loses too much accuracy on the *positive-low* class in order to improve average accuracy. However, no strategy does generate drastically worse results than the best modality - actually they perform well compared to remaining modalities. This is a good indication that introduced strategies to handle missing data guarantee a robust recognition result.

So in order to perceptibly enhance recognition rates compared to single channel classification on the dominant modality, we have to exploit deeper knowledge about the classification problem at hand with emotion-adapted fusion strategies that employ more than a single generic ensemble (Table 9.4). The combination of arousal and valence ensembles shows different characteristics than the generic approaches: The dominance of the *positive-high* class is gone and negative classes are well recognised. Overall these changes result in slightly superior accuracy than the best single channel. For further improvements we incorporate more available information from the 2D emotion model into combination strategies, leading to the additional cross axis ensemble. This fusion

Emotion-adapted Decision Level Fusion		
	Voting without CrossAxis	**Voting with CrossAxis**
Positive-Low	56.0%	64.0%
Positive-High	46.0%	55.0%
Negative-Low	55.0%	56.0%
Negative-High	51.0%	44.0%
UA	52.0%	55.0%

Table 9.4: Results of emotion-adapted fusion approaches are in line with earlier positive findings and show that a positive multi-modal effect can be achieved even in a scenario with missing data.

scheme exceeds the best modality on every observed class and therefore enhances average accuracy remarkably, however at the expense of a rising ensemble count. These results are in line with positive findings about emotion-adapted fusion approaches (Chapter 5.2) and show that a positive multi-modal effect can be achieved even in a scenario with missing data.

9.2 Event-driven Fusion in the Wild

Asynchronous fusion on event level has proven to be robust in affect recognition scenarios (Chapter 8) and provides an abstraction level that allows to build a highly adaptable recognition systems, as modalities that contain information about a sought target class and provide events for the fusion algorithm can (from a technical point of view) be easily added or removed. Therefore it is a good fit for in the wild signal processing, where there is no guarantee to have all sensors available at all times.

Figure 9.3 shows the theoretical behaviour of event-driven fusion results when all modalities are available (Figure 9.3 (1)) or when a subset of the possible input events is missing (Figure 9.3 (2)). In a scenario where we expect multiple events to be active during interesting time segments, the omission of several events could lead to a less unambiguous but still acceptable assertion. In order to evaluate this assumption, we decided to perform an affect recognition task in an everyday setting. In Chapter 8 we developed an event-driven fusion system to recognize user enjoyment based on audiovisual cues in the form of laughs and smiles and we will now build upon this system to realize a mobile scenario.

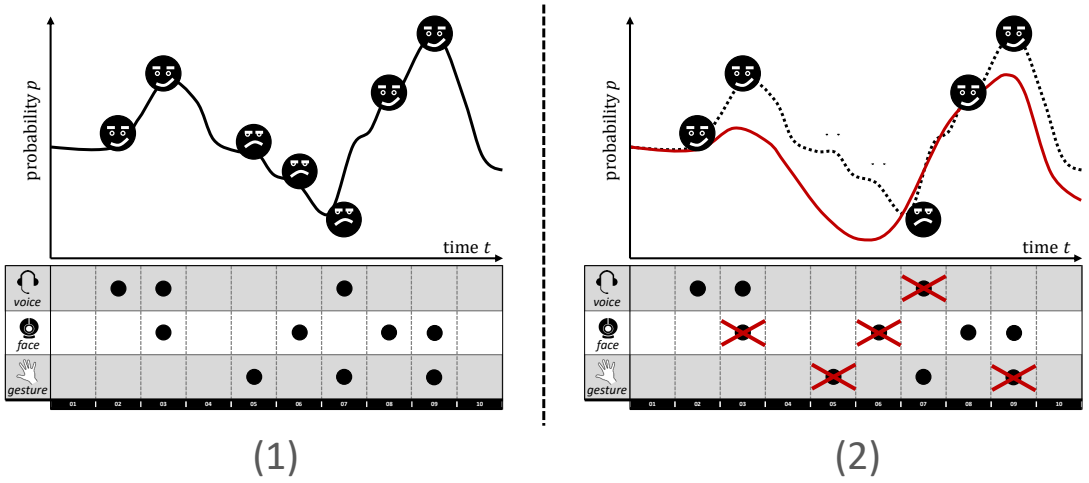


Figure 9.3: On the left side (1) we see the theoretical behaviour of an event-driven fusion algorithm with all modalities consistently available. On the right side (2) 50% of events are unavailable. The omission of several events leads to a less unambiguous but still acceptable assertion.

9.2.1 Mobile and Multi-modal Laughter Recognition

In the evaluation in Chapter 8, data acquisition was done in a typical stationary lab setting in which up to four study participants were recorded in group converstion, telling amusing stories from personal experience (McKeown *et al.* [117]). The topics of the conversation were guided by an external expert. Our aim towards multi-modal affect recognition in the wild is consequently to reinterpret this setting in a natural environment, without external guidance and with a more unobtrusive technical set-up (Flutura *et al.* [63]). To do so, it would be optimal if all needed sensory equipment could be provided by a hand-held device, such as a mobile phone. Given these technical restriction and the general problems of a real-life setting and an unguided conversation, we ease the recognition problem: Instead of recognizing the enjoyment level of probands, we concentrate on the multi-modal recognition of laughter.

Regarding available modalities, they are strongly dictated by the natural setting. Audio is a main source of information regarding the recognition of laughs, and by using clip-on microphones we can achieve personal recordings of single users. The visual modality, that was included in previous experiments is however problematic: A frontal camera capturing a user's face to detect visual laughter is hard to realize in a real-life setting. Placing cameras in the environment is no option in a potentially mobile setting as they restrict the possible movement. Other ways of attaching cameras to the user himself would result in hindering constructs. A possible alternative offered by mobile devices is accelerometer data. Related recognition systems have successfully used motion data

to recognize laughter. Mancini *et al.* [113] have used visual markers to track head and shoulder movement in order to calculate a body laughter index. McKeown *et al.* [116] systematically investigated body movements during laughter episodes using whole body motion capturing. Consequently we will try to replace the Kinect™device used in the laboratory setting with accelerometers offered by mobile devices.

Figure 9.4: MobileSSI brings social signal processing to mobile devices. It provides a flexible framework for synchronously interacting with multiple wearable sensing devices in real-time while not constricting the user's mobility. The depicted deployment in a natural pub setting demonstrates its capability to run complex signal processing and machine learning tasks locally on mobile devices. In this case users are equipped with smartphones and clip-on microphones to enable multi-modal laughter recognition in the wild.

9.2.2 Evaluation in a Real-life Scenario

As evaluation scenario we decided to stick to the interpersonal conversation setting. However, instead of having the probands invited into a fixed laboratory set-up, they are exposed to an actual pub as a realistic environment for enjoyable conversations. Underlining the naturalness of the setting, there is no external guidance on chosen topics. Compared to the Belfast Storytelling Database (Chapter 8.1), we can expect a drop in signal quality: In the pub scenario the audio signal is overlaid with environmental noise that is not given in a laboratory set-up, such as background music or surrounding conversations of varying intensity. These disturbances have negative influence on event

detection and audio classification and consequently lead to missing events and wrong interpretations.

Regarding the recording within a natural and mobile setting, we chose to port the Social Signal Interpretation framework (SSI) (Wagner *et al.* [169]) to run on Android™based mobile phones. This way we are able to provide a flexible framework for recording and real-time interpretation of multiple wearable sensing devices without constricting the user's mobility. MobileSSI is open source and available to the public[2]. For the analysed recordings, users were equipped with a smartphone (including accelerometer sensor) and a connected clip microphone (Figure 9.4).

As in former studies, we use 1451 paralinguistic features offered by EmoVoice (Vogt *et al.* [166]) to process the audio modality. The feature set for accelerometer data consists of a series of nine features for each of the three accelerometer axis (Figure 9.5) together with their first and second derivatives, resulting in 81 accelerometer features.

modality	**short-term feature**	**long-term feature**	**total**
accelerometer	3 axes acceleration, 1st and 2nd derivatives	mean, standard deviation, minimum/maximum, range, zero crossing rate, peak count, pulse rate, energy	81

Table 9.5: Overview of feature extraction methods applied to the accelerometer modality.

Evaluation is carried out over the two recorded pub sessions. As in Chapter 8 we implement a framewise and user-independent approach. Both sessions are recorded with the same users and we hold recordings of a single user back from the training process. Table 9.6 shows unweighted recognition results gained by uni-modal support vector machine classification of laughter frames via the audio and accelerometer channels.

	Uni-Modal Classification	
	Accelerometer	**Audio**
Laugh	80.95 %	76.19 %
¬ Laugh	63.42 %	86.70 %
Average	**72.19 %**	**81.45 %**

Table 9.6: Results of uni-modal laughter classification.

Compared to laughter recognition from the audio modality in the laboratory setting (Chapter 8.3.2), we can observe a drop from 84.05% to 81.45% in recognition accuracy.

[2] `https://hcmlab.github.io/mobileSSI/`

Given the difficult noise situation in the pub scenario, this drop is expected. As indicated before, visual laughter detection was excluded in the mobile scenario and replaced with laughter recognition from accelerometer data. While the video modality is a very common and reliable source for laughter recognition, the implementation of movement features to detect hints for laughing are more seldom and expected to be less reliable. Therefore it is no surprise, that a direct accuracy comparison of the visual laughter detection in the laboratory setting with movement based laughter detection on the pub environment shows a negative difference of 6.72% (72.19% (accelerometer) compared to 78.98% (video)).

	Multi-Modal Classification	
	Decision Fusion	**Event Fusion**
Laugh	78.57 %	83.33 %
¬ Laugh	86.62 %	85.95 %
Average	**82.59 %**	**84.64 %**

Table 9.7: Results of multi-modal fusion on decision (product rule) and event level (gravity fusion).

In Table 9.7 we compare a basic decision level fusion approach (product rule) to the event-driven fusion strategy. Synchronous combination of the audio and accelerometer modality lead to a slight performance gain compared to using the audio modality only. Asynchronous fusion of audio and movement cues on event level improves classification by 3.19%. Instead of fusing information over fixed time segments, recognized events are integrated frame by frame by the gravity fusion algorithm (Chapter 7.3.2). Hereby, we learned the optimal parameters for initial event weights and decay speed via grid learning over possible combinations with the 18 best configurations averaging around 84%. These combinations advise to weight the audio modality higher than movement cues, which is plausible considering the difference in single modality performance (Table 9.6). Audible laughter cues are also given a longer lasting influence on the fusion result than the more short-lived movement cues.

9.3 Conclusions and Further Challenges

In this chapter we have shown how the challenges of affect recognition in the wild - especially the problem of missing data - can conveniently be tackled within the multi-modal fusion step. Synchronous fusion approaches generally expect input from all included ensemble members, so in order to handle unavailable input, enhancements of chosen algorithms have been suggested and evaluated. We see that generic fusion approaches adapt to the most reliable modality in the ensemble and while they are able to compensate the temporal loss of information, they are not able to generate a positive multi-modal effect. This behaviour is in line with findings about the assumed interchangeability of these fusion strategies (Chapter 6). If we however exploit the structure of the underlying emotion model with emotion-adapted fusion schemes, there is the possibility to make better use of the multi-modal information and gain recognition enhancements even with missing data. These fusion schemes are able to outperform single modalities and generic approaches by a significant rate, though bearing a higher complexity due to the generation of specialised ensembles.

The inherent logic of an event-driven fusion approach should basically be better suited to cope with partially missing, unreliable or noisy data than synchronous approaches. As events are introduced as an abstract intermediate layer that effectively decouples uni-modal processing from the final decision making, missing input from one of the modalities does not cause the collapse of the whole fusion process. Embedding this fusion logic into the MobileSSI framework, which is able to run signal processing and machine learning techniques on mobile devices, we were able to realize a multi-modal laughter recognition application in a natural social setting and demonstrate the successful transition from a laboratory setting to the wild. The video modality is classically well suited for laughter recognition, but it is not suitable in the proposed mobile setting. We have consequently replaced it within the fusion process with accelerometer based recognition. Although uni-modal classification results are lower than an analogous evaluation in a controlled environment obtained in the lab (Chapter 7), techniques based on event fusion lift the laughter recognition accuracy with audio and accelerometer events on par with laughter recognition from the audio channel in a laboratory environment.

Challenges that emerge in natural and mobile scenarios of course go beyond the problem of missing data and cannot be solved by an appropriate fusion technique alone. In the mobile setting described in Section 9.2 we have seen that environments are subject to great changes, e.g. when the users switch between indoor and outdoor locations (Figure 9.5). Noise cancellation schemes are required that are able to dynamically adapt to

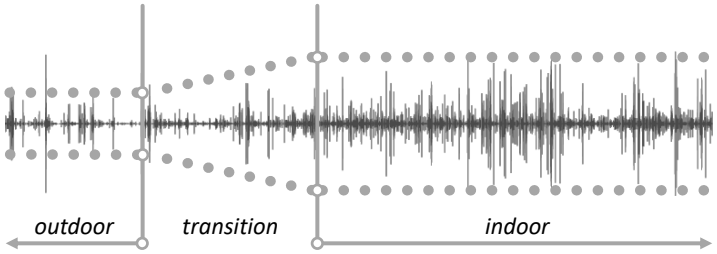


Figure 9.5: The challenge of changing environments needs to be tackled in mobile settings. An example are differing noise conditions in the audio modality, appearing during the transition from an outdoor to an indoor scenario.

the current situation. Another approach to handle this problem would be the automatic categorization of the current surrounding and the use of pre-tailored classification models outdoor and indoor settings. MobileSSI has shown to be a well suited tool for affect recognition outside the laboratory. Since battery life of modern mobile devices is sufficient to record and process data in real-time for several hours, we will be able to run more real-life experiments, which provide better insights on the actual challenges we have to face when applying social signal processing in the wild.

Chapter 10

Multi-modal Fusion in Applications

In the preceding chapter we have seen the practical realisation of a multi-modal fusion system for a real-world application. In the following, we will go into detail about all steps needed for the concrete implementation of multi-channel fusion systems. All approaches that have been theoretically discussed and evaluated in this thesis are included as re-usable components in the Social Signal Interpretation framework, which supports synchronized and (near) real-time processing of multi-sensor data as well as subsequent machine learning tasks. These features lays the perfect foundation for the realisation of multi-modal affect recognition pipelines and we will explain the needed steps to implement a complete pipeline from sensor input to a multi-modal emotion recognition result. Afterwards we will have a look at several fusion systems that have been successfully deployed as affect recognition components within Human Computer interaction (HCI) systems implemented in the course of European research projects.

10.1 Implementation in the Social Signal Interpretation Framework (SSI)

The Social Signal Interpretation (SSI) framework (Wagner *et al.* [170]) offers an open source solution to realize machine learning pipelines from sensory input to final affect recognition. SSI is written in C++ and optimized to run on computer systems with multiple CPUs. Binaries and source code are freely available[1] under GPL. Its main strength lies within the patch based design that allows to create applications from

[1] `http://hcm-lab.de/projects/ssi/`

existing modular components. Hereby we can choose to include any desired amount of input modalities which are recorded and processed in a synchronized manner. This trait is the foundation to apply multi-modal fusion. Consequently, all presented fusion techniques and their evaluation we have seen in previous chapters as well as practical fusion applications presented in this chapter have been implemented as SSI components and we will at this point include some introductory examples how different fusion approaches can be easily realized as SSI pipelines.

10.1.1 Basic Concepts and Workflow of the Framework

The core concepts needed to establish a multi-modal affect recognition pipeline within the SSI framework can be achieved by explaining a few basic components (Figure 10.1). The whole range of the functionalities offered by the framework is out of scope of this thesis, however, a detailed documentation is of course available[2].

- **Sensors and synchronized data streams**
  Each and every affect recognition process starts with input obtained from convenient hardware devices. To capture relevant emotional input from affective channels, a diverse list of sensory hardware needs to be incorporated. The signals provided by these *sensors* (and all successive components in the pipeline) are encapsulated in a common data structure - the data *stream*. By making sure all connected sensors start providing data at exactly the same time and monitoring the adherence of specified sample rates, SSI guarantees synchrony among data streams within the system.

- **Signal processing with transformer components**
  For every affective channels within the system, we need ways to pre-process the raw data (e.g. filtering or normalization) and extract expressive features from the signal streams. Respective steps are realized as so called *transformers*, a component interface designed to take one or more data streams as input. This input is processed in adequate segments and the transformed information is again provided as data stream. This way the data can be processed further by subsequent components in the framework.

[2] `https://myweb.rz.uni-augsburg.de/~wagjohan/openssi/`

- **Classification and fusion in consumer components**
  After all needed processing steps are applied and meaningful features are calculated, the extracted information can be used for classification and multi-modal fusion. The designated component type to achieve this task is the *consumer*. Regarding the flow of data streams, consumer components are an ending point. They accept data streams as input, but no output in the form of data streams is produced. Instead the received information is evaluated, for example, by classification algorithms. As a consumer is able to receive several input streams in parallel and is able to hold multiple classification models, it a convenient component to implement synchronous fusion strategies.

- **Asynchronous communication between objects via events**
  All components within the framework are derived from the general *object* interface, which is able to generate and receive timed *events*. This approach enables the asynchronous communication between components and can be used to implement an event-driven fusion approach. To give an example we can define several classification components within a pipeline that propagate found affective cues as events to a centralized event board. An event fusion component (object) registers itself to the event board as receiver of these events, monitors and evaluates their occurrences and fuses them into a multi-modal event-driven decision.

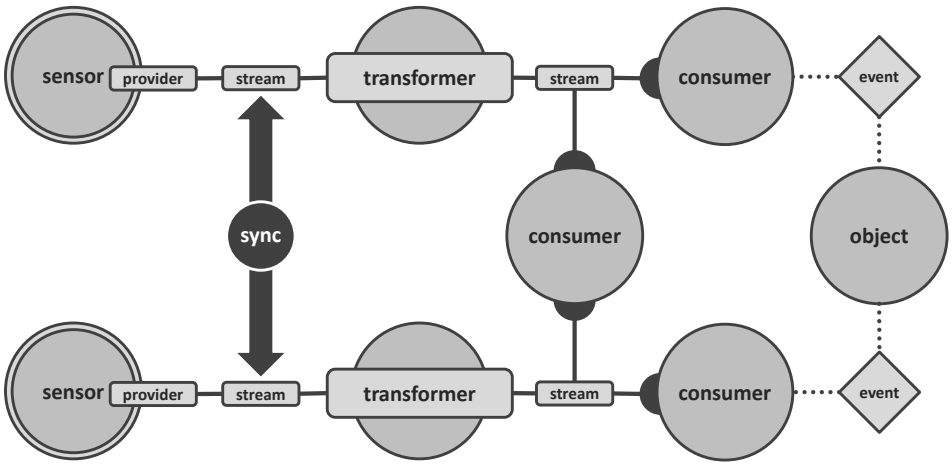


Figure 10.1: Signals provided by *sensors* are encapsulated in the general data structure of a data *stream*. Signal processing and feature extraction steps are realized in *transformer* components. *Consumer* components are the end point of pipelines - as far as signal streams are concerned. Classification and synchronous fusion can be realized here. All components are derived from an *object* interface, which is able to generate and receive timed *events*. This enables the asynchronous communication between components and can be used to implement an event-driven fusion approach.

Figure 10.2 shows the exemplary workflow of the SSI framework. Within a real-time pipeline we access the sensors that are e.g. meant to monitor the voice and the facial expressions of a user. Generated data streams can now be processed online as well stored on disk for later offline analysis. The processing includes data pre-processing and filtering, activity detection and feature extraction. These steps can be applied to online data streams or stored data. In the offline case, the processed data - in conjunction with annotations - can be used to train and evaluate classification models, which are later installed in real-time recognition pipelines. Given the fact that the framework is able to handle several data streams in parallel, we have the ability to include multi-modal fusion (normally) at the end of a recognition chain.

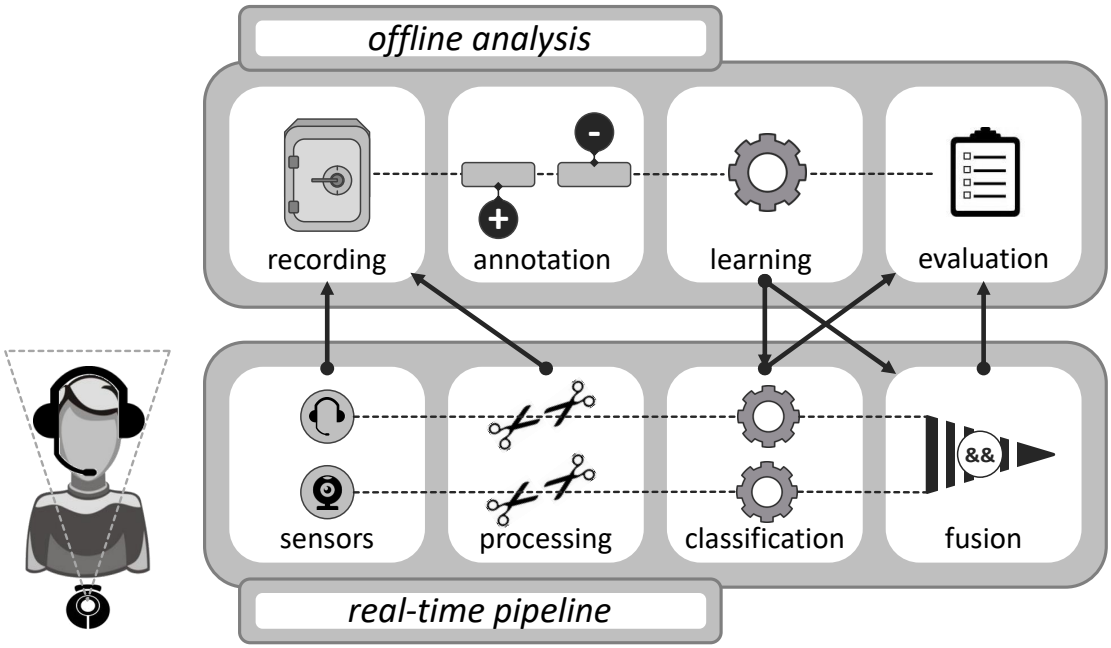


Figure 10.2: Schematic workflow of the SSI framework. Sensor data can be stored on disk and together with annotation tracks be used to train and evaluate classification and fusion models. On the other hand the data can be processed in real-time and fed into pre-trained classifiers. The parallel handling of multiple input sources allows the implementation of multi-modal fusion systems.

10.1.2 Example Pipelines

After explanation of the basic components and the theoretical workflow of the SSI framework, we are now ready to go through a bottom-up example for affect recognition. We start with uni-modal classification, afterwards add a second modality and perform a standard fusion algorithm. Lastly we make use of the inherent event logic of the framework to introduce an event-driven fusion approach to the pipeline.

To help users without deeper programming experience get started with pipeline construction, SSI offers the possibility to define the processing chain with XML files. At runtime, the XML specifications are translated into pre-complied C++ code. Please note that though this approach is very convenient to design any signal processing pipeline with existing components, there is no way to implement new components with XML.

Uni-modal Classification

We start with the most simple way to achieve an affective classification from an affective channel - the uni-modal classification (Figure 10.3). We choose the audio signal as our first modality, extract paralinguistic features and classify the affective state with a pre-trained classification model.

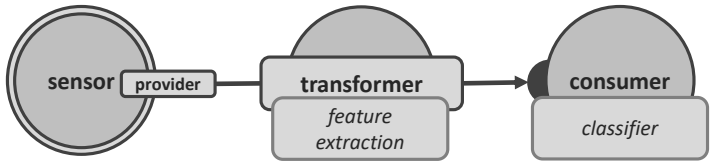


Figure 10.3: Uni-modal classification using a single sensor, a feature extraction component and a pre-trained classifier.

The XML pipeline starts with the introduction of an audio sensor `<ssi_sensor_Audio>`. The entry `<pin="stream_audio">` defines the data stream which provides the raw audio data. The next component in the chain in the feature extraction component in form of a transformer.

As in preceding chapters, we use the statistical EmoVoice features (Vogt *et al.* [166]) `<ssi_feature_EmoVoiceFeat>` to extract emotional content from the raw audio data. The entry `<input pin="stream_audio" frame="1.0s">` tells the transformer to use the specified data stream as input, processing segments containing one second of data. Calculated features are again provided in a stream format, defined by `<output pin="stream_audio_feat">`.

```
<sensor create="ssi_sensor_Audio" sr="48000" scale="true">
 <provider channel="audio" pin="stream_audio"/>
</sensor>

<transformer create="ssi_feature_EmoVoiceFeat">
 <input pin="stream_audio" frame="1.0s"/>
 <output pin="stream_audio_feat"/>
</transformer>

<consumer create="ssi_consumer_Classifier" trainer="trainer_audio">
 <input pin="stream_audio_feat" frame="1"/>
</consumer>
```

Classification is carried out in the consumer component <ssi_consumer_Classifier>. It takes the feature stream as input and generates the decision with a pre-trained classification model. Among other information, the location of this classification model on disk is provided in a separate trainer file <trainer="trainer_audio">. The entry <frame="1"> tells the consumer to process every single data segment, in this case one segment corresponds to one feature vector over one second of raw data.

The <trainer_audio> file contains information about the shape of the expected stream, available classes to be classified and the location of the classifier - in this case a support vector machine <ssi_model_SVM>.

```
<trainer>
 <info trained="true"/>
 <streams>
  <item byte="4" dim="1451" sr="1.000000" type="FLOAT"/>
 </streams>
 <classes>
  <item name="happy"/>
  <item name="unhappy"/>
 </classes>
 <users>
  <item name="user"/>
 </users>
 <model create="ssi_model_SVM" stream="0"
  path="trainer_audio.trainer.SVM"/>
</trainer>
```

Standard Fusion

Next, we make the step to multi-modal fusion (Figure 10.4). Therefore we introduce a second modality, action units provided by a Microsoft Kinect™ sensor `<ssi_sensor_MicrosoftKinect>`.

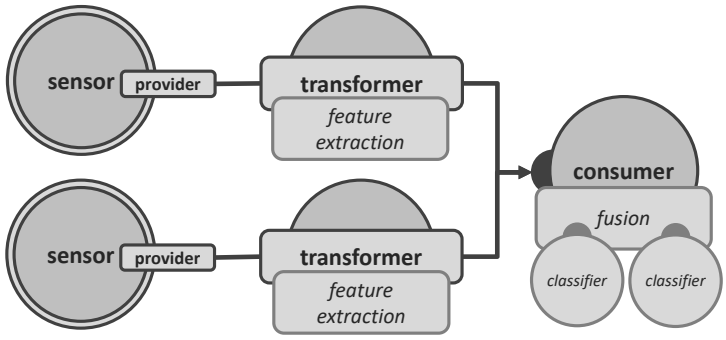


Figure 10.4: Decision level fusion using a second modality and a consumer taking two streams containing features as input.

```
<sensor create="ssi_sensor_Audio" sr="48000" scale="true">
 <provider channel="audio" pin="stream_audio_raw"/>
</sensor>

<sensor create="ssi_sensor_MicrosoftKinect" sr="25">
 <provider channel="kinect_au" pin="stream_kinect_au"/>
</sensor>

<transformer create="ssi_feature_EmoVoiceFeat">
 <input pin="stream_audio" frame="1.0s"/>
 <output pin="stream_audio_feat"/>
</transformer>

<transformer create="ssi_feature_MicrosoftKinectAUFeat">
 <input pin="stream_kinect_au" frame="1.0s"/>
 <output pin="stream_kinect_au_feat"/>
</transformer>

<consumer create="ssi_consumer_Classifier" trainer="trainer_fusion">
 <input pin="stream_audio_feat" frame="1"/>
 <xinput size="1">
  <input pin="stream_kinect_au_feat"/>
 </xinput>
</consumer>
```

Analogous to the audio signal, we calculate statistical features <ssi_feature_MicrosoftKinectAUFeat> for segments of one second Kinect action unit data. Fusion is carried out in a consumer component, that at first sight resembles the uni-modal classifier. By closer inspection, we see that the <ssi_consumer_Classifier> expects an additional stream <xinput>. This of course is needed for the fusion step, where the information from both modalities is combined. In this case we use a decision level approach as fusion strategy. This is reflected in the <trainer_fusion> file:

```
<trainer>
 <info trained="true"/>
 <streams>
  <item byte="4" dim="1451" sr="1.000000" type="FLOAT"/>
  <item byte="4" dim="36" sr="1.000000" type="FLOAT"/>
 </streams>
 <classes>
  <item name="happy"/>
  <item name="unhappy"/>
 </classes>
 <users>
  <item name="user"/>
 </users>
 <fusion create="ssi_fusion_ProductRule"
  path="trainer_fusion.trainer.ProductRule">
  <models>
   <item create="ssi_model_SVM" stream="0"
    path="trainer_fusion.trainer.#0.SVM"/>
   <item create="ssi_model_SVM" stream="1"
    path="trainer_fusion.trainer.#1.SVM"/>
  </models>
 </fusion>
</trainer>
```

Compared to the <trainer_audio> file we now see definitions for two expected streams and the two respective classification models are provided. Furthermore the product rule <ssi_fusion_ProductRule> is defined as combination strategy for the two classifier decisions. Eventual fusion parameters such as modality weights (e.g. for weighted voting approaches) or additional classification models (e.g. for score level fusion) are provided in the <trainer_fusion.trainer.ProductRule> file, which generated during the offline training process.

Event-driven Fusion

The just explained method of fusing multiple modalities within a single consumer allows asynchronous fusion approaches by including classification models with memory capability. However, to implement an event-driven fusion strategy we need to include SSI's event logic in the pipeline. To add this feature, we keep the audio and Kinect sensors as well as the respective feature extraction components in place, but replace the fusion consumer with the following uni-modal classification consumers:

```
<consumer create="ssi_consumer_Classifier" trainer="trainer_audio"
          address="laugh@voice">
 <input pin="stream_audio_feat" frame="1"/>
</consumer>

<consumer create="ssi_consumer_Classifier" trainer="trainer_kinect"
          address="smile@face">
 <input pin="stream_kinect_au_feat" frame="1"/>
</consumer>
```

The addition of an <address> to a classification consumer enables the distribution of its results as events through an event board. The structure of the event-address is given by the identifying name of the event and the name of the sending component, connected by the '@' char. Any component is now able to receive the specified events and process them (Figure 10.5) - exactly the structure needed for the realisation of event-driven fusion.

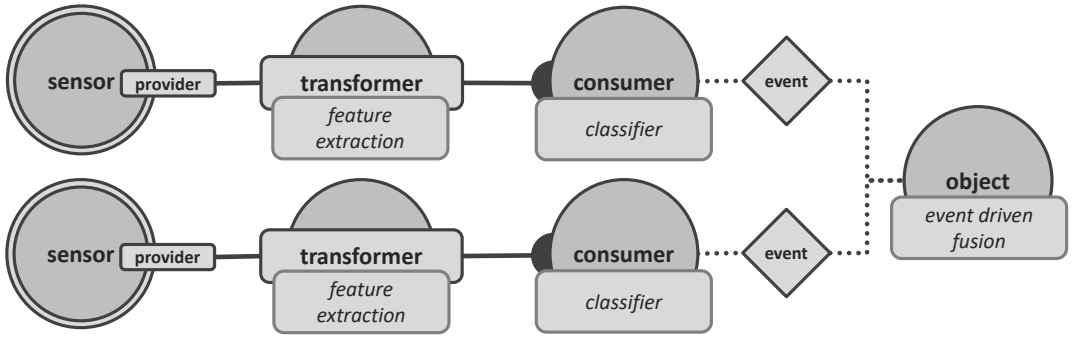


Figure 10.5: Simple event-driven fusion schematic. Classification consumers propagate their results as events that are handled by the fusion component.

As an event-driven fusion component will typically not need direct stream input and only work with received events, it is implemented as an SSI object. In this example we insert

a <VectorFusionGravity> object, which listens to the affective events recognized in the audio and facial modalities. On basis of these events, the fusion result is calculated and can again be propagated to the application.

```
<object create="VectorFusionGravity"
        dimension="2" fusionspeed="1.0f"
        threshold="0.1f" update_ms="100"
        path="parameters.modality"
        address="event@fusion">
 <listen address="laugh,smile@voice,face"/>
</object>
```

The <parameters.modality> file defined in the fusion component hold the parameters such as speed and weight that can be defined per event.

```
laugh@audio    0.05   0.5
smile@video    0.10   1.0
```

Of course this example shows a rather basic event-driven fusion pipeline. For the sake of tutorial purposes we excluded pre-processing of raw data streams. Furthermore, at this point affective events would be generated in each modality for each successively analysed frame. Activity detection such as voice activity detection in the audio channel and eventual normalisation and mapping steps for recognised events would need to be included.

Equipped with the background knowledge how to implement multi-modal affect recognition pipelines we can at this point go into concrete applications implemented in the course of several European research projects.

10.2 The CEEDs Project - Data Level Fusion with Physiological Responses

The European research project CEEDs (Collective Experience of Empathic Data systems)[3] aimed at developing an intelligent environment that helps a user handle the eventually overwhelming confrontation with *Big Data*. As an elaborate use case, a scientific discipline that traditionally produces huge and complex datasets was chosen: Neuroscientific

[3] https://ceeds-project.eu/

data is processed to be animated with visual and sound stimuli within a 3D chamber (XIM - eXperience Induction Machine) with the aim to make it more accessible to the user (Figure 10.6). The idea behind this immersive scenario is to build a system that monitors implicit user input and guides the experience on a subconscious level, e.g. by fluidly altering the complexity of presentations or providing subliminal guidance. Consequently we use continuous measurements of physiological user responses to influence the presentation of the data with respect to optimal complexity and immersiveness.

Figure 10.6: The CEEDs eXperience Induction Machine (XIM) presents huge neuroscientific datasets with visual and sound stimuli. Subconscious user responses are used to guide the experience with regard to optimal complexity and immersiveness.

To make these subconscious user reactions available to the XIM, we need unobtrusive physiological sensory devices and real-time analysis of multi-modal data. To this end a sensing glove, a sensing shirt and a wearable eye camera were developed, providing galvanic skin response (GSR), electrocardiography signal (ECG) and resulting heart rate as well as the current pupil dilation (PD). These signals serve as input to the multi-modal fusion system (Figure 10.7) that is meant to assess current levels or arousal and cognitive workload.

From the viewpoint of an eligible fusion strategy, we are at this point confronted with the challenge that the project did not foresee the recording of appropriate training data fitting these unconventional conditions. Consequently, supervised learning is not a viable option. Therefore a data fusion approach (Chapter 4.1.1) - a rather unconventional fusion strategy for affect recognition tasks - is applied: The heart rate calculated from the ECG signal and the phasic component of the GSR signal are normalized and merged into a single signal. This is done with the prior knowledge that both modalities feature the characteristic to increase in amplitude and peak frequency during episodes of high

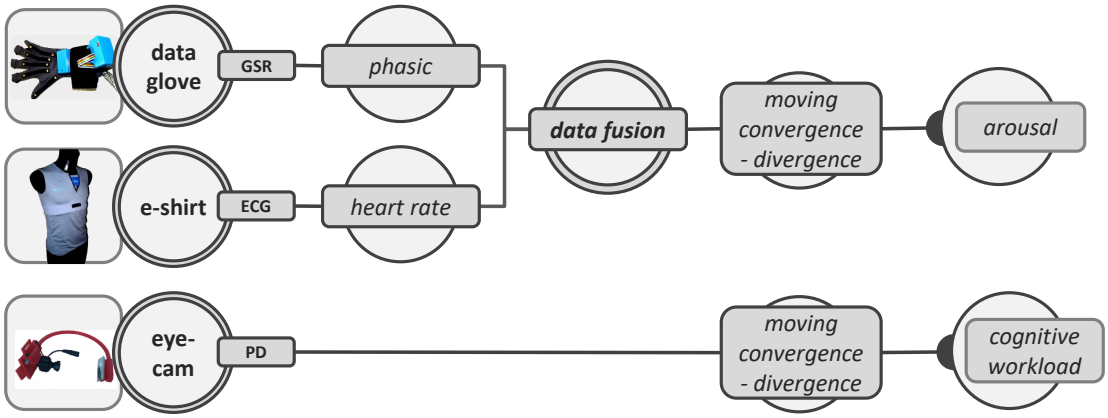


Figure 10.7: CEEDs user response recognition system. A data fusion approach is applied to combine the normalised heart rate signal with the phasic component of galvanic skin response to assess the arousal level of a user.

arousal. These features should consequently be observable in the fused signal. On the one hand, they should be visible in the fused signal even if these features only emerge in one modality, on the other hand they could even be reinforced in the fused signal when emerging in both modalities at about the same time.

To achieve workable assessment for the whole system, we apply a moving average convergence-divergence (MACD) filter to the fused signal. The MACD filter is a trend-following momentum indicator that expresses the changes between short and long term moving averages of a signal and is traditionally used for the technical analysis of stock prices. We apply this calculation to the fused heart rate and phasic galvanic skin response signal and use it to detect rising and falling levels of physiological user arousal. The same momentum is calculated directly on the signal describing pupil dilation, as this measurement can be directly related to cognitive workload (Chapter 2.2.4).

10.3 The ILHAIRE Project - Event-driven Enjoyment Recognition in Human-Avatar Communication

Laughter is a significant feature of human communication and the European research project ILHAIRE (Incorporating Laughter into Human Avatar Interactions: Research and Experiments)[4] concentrated on the the gap between knowledge on human laughter and its use by avatars. This included the multi-modal recognition as well as synthesis of

[4] http://www.ilhaire.eu/

laughter to drive the implementation of conversational agents with the social capability to spot and use laughter in natural human-avatar interaction scenarios (Figure 10.8).

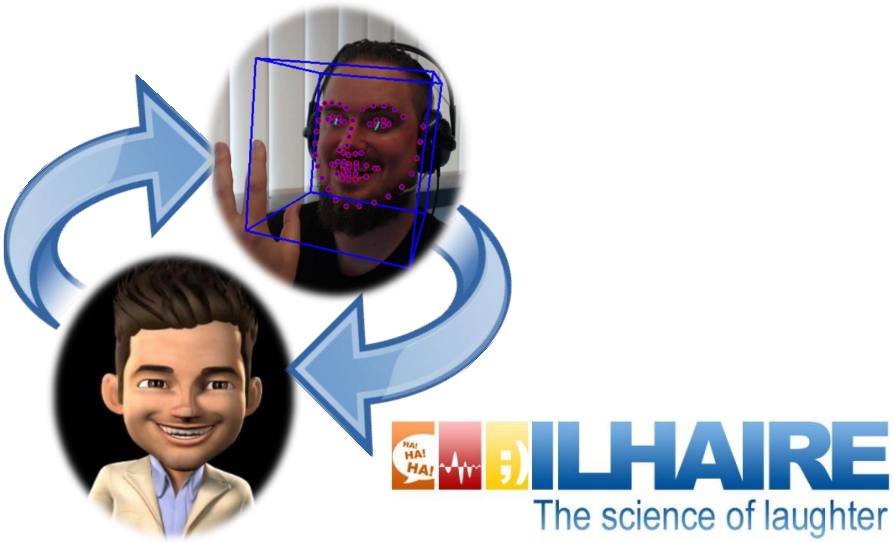


Figure 10.8: The reliable recognition and high quality synthesis of laughter are great features to enhance the naturalness of human-avatar interaction.

Our initial main focus was on the robust recognition of human laughter in conversations. Evaluations in Chapter 8 show that laughter events can reliably be detected in the vocal modality. With this result at hand and a natural human-avatar interaction as goal, we came to the conclusion that the recognition of longer-lasting enjoyment episodes could be more usable in interaction design than the short-termed bursts of laugh. To achieve a continuous assessment of a user's current enjoyment level, we developed the concept of an event-driven fusion approach: As discussed in Chapter7 we identified enjoyment as a rather abstract target class that can however be well characterized by the cumulated occurrence of laughs and smiles. Laugh and smile events are generated by an activity check for each modality (face tracked and frontal, voice activity) for each frame and a succeeding binary SVM classification for laugh and no-laugh as well as smile and no-smile respectively. The resulting classifier probabilities are used as event scores. The fused enjoyment score can therefore on the hand be interpreted as confidence value for current enjoyment, on the other hand we see in practical application that this score often also correlates with the perceived intensity of enjoyment.

In addition to audiovisual modalities the project also investigated the capabilities of more experimental approaches such as shoulder movements and respiration patterns. Figure 10.9 shows the event-driven enjoyment recognition system with maximal number of considered modalities. In the final evaluation (Chapter 8) only audiovisual channels were incorporated, as the experimental sources for laughter cues showed promising initial results but also flaws in technical realisations: The shoulder tracking was based on coloured markers attached to the shoulders of a user. This setup is rather obtrusive and unnatural and a colour based tracking approach also showed difficulties. The obtrusiveness argument also holds for the respiration belt whose recognition accuracy did finally not make up for the drawbacks.

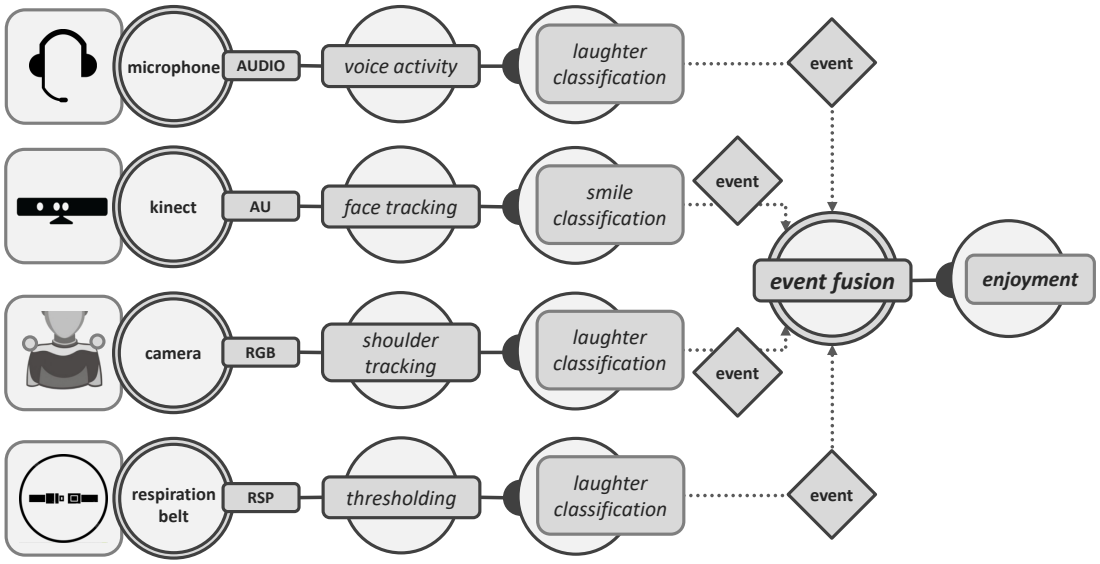


Figure 10.9: The event-driven enjoyment recognition system of the ILHAIRE project, with maximal number of considered modalities. Apart from robust audiovisual event detection, we considered additional experimental modalities as sources for enjoyment related cues.

A more unobtrusive and practicable approach to spot laughter cues from body movements is based on accelerometer data of mobile devices and has been presented in the mobile laughter recognition system in Chapter 9.

10.4 The KRISTINA Project - Event-driven Fusion in the Valence-Arousal Space under Changing Conditions

The currently ongoing KRISTINA project[5] (A Knowledge-Based Information Agent with Social Competence and Human Interaction Capabilities (Wanner *et al.* [173])) is the last EU funded project to be presented in this series of research ventures. The goal is to develop a human-like socially competent and multilingual avatar meant to deal with problems within the health domain. The special focus hereby lies on problems regarding language and cultural barriers that migrant patients and representatives of the host country's health system may face. The avatar is among other tasks given the role of a first information source for migrants regarding healthcare issues and can also be applied as a mediator between migrant patients and local caregivers. In addition to the challenge of dealing with multiple languages, the avatar is meant to represent a trustworthy contact partner and therefore a natural interaction in which emotional expressions are considered is a must (Wanner *et al.* [174])(Figure 10.10).

The consequently applied multi-modal emotion recognition system is again an event-driven approach. In contrast to the before mentioned fusion systems, we do not deal with a single target class but with an emotional assessment within the whole valence-arousal space. If we look at the schematic of the fusion system (Figure 10.11), we see that the abstraction step from information in affective channels to events used in the event-driven fusion algorithm is handled by the uni-modal recognition components. These provide affective cues already translated in valence and arousal values. In detail, we classify continuous valence and arousal scores from the audio modality via recurrent neural networks (i.e. LSTM-NN), derive respective scores from facial action units with linear regression and calculate an arousal score from movement vectors of the hands. This classification is carried out for every frame (the frame size of course varies per modality) after respective activity check (hands tracked, face tracked, voice activity) and the results are published as events.

The fusion algorithm handles the incoming valence and arousal events with respect to their occurrence over time. The targeted valence-arousal space spans two axes. The event-driven fusion strategies presented in Chapter 7 are of course able to handle multidimensional events, so the processing of events holding two entries (one current valence and one current arousal score) is theoretically possible. However, this approach

[5] `http://kristina-project.eu`

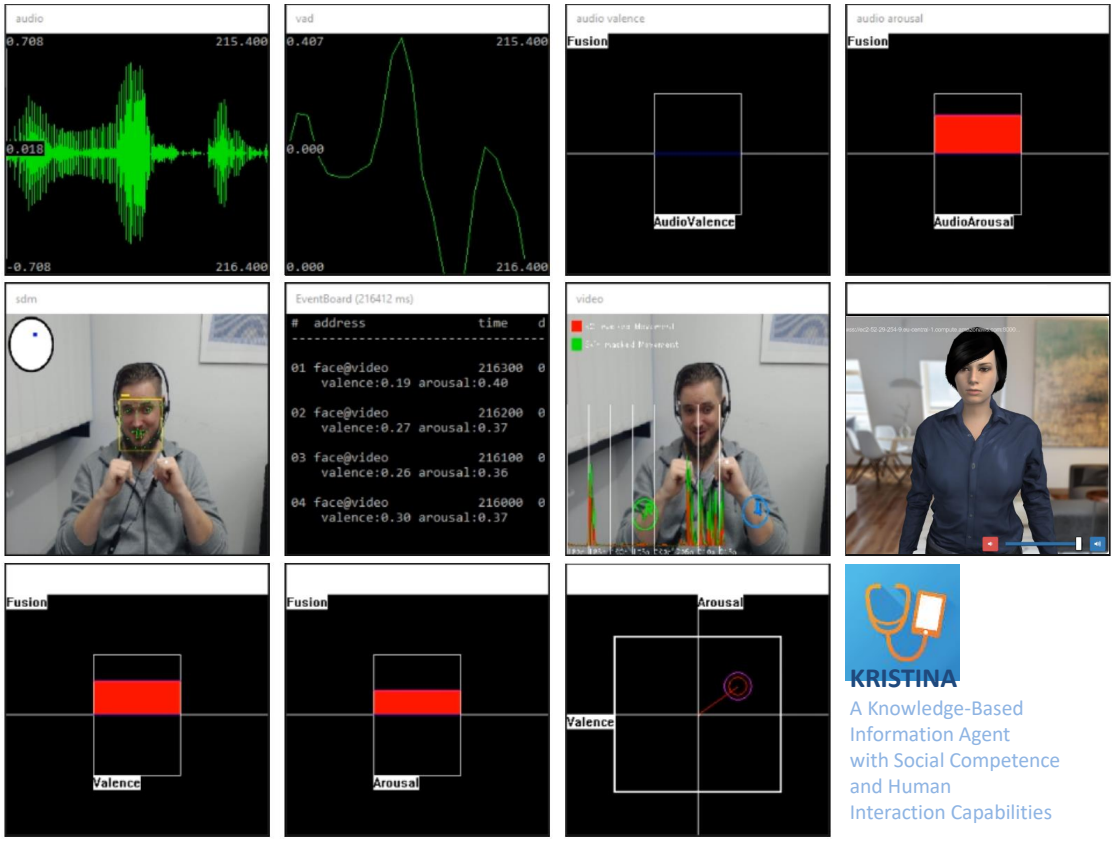


Figure 10.10: The multilingual KRISTINA avatar wants to represent a trustworthy contact partner and therefore needs to offer a natural interaction in which emotional expressions of the user are considered.

induces an interdependence between the two axes, which may be beneficial if the dimensions of a recognition problem for example mutually exclude each other (e.g. happy - unhappy). In the valence-arousal space this behaviour is not desirable, the two dimensions are generally independent. The system schematic (Figure 10.11) also shows that the gesture modality only provides reasonable insights in the arousal level. If we would demand two-dimensional events, the valence entry would need to be filled, but every possible entry (e.g. zero) would suggest an unwanted influence on the fusion result. The solution is the definition of two separate event processing fusion components, one for each axis of the valence-arousal space. The single results are merged back into a two-dimensional final result to be published to the avatar application.

Another challenge given by the application are the changing environmental conditions under which the KRISTINA avatar should be able to be addressed. It should on the one hand be used as a traditional desktop application, as well as a mobile application on hand-held devices. For modalities available to the affect recognition system this means a changing availability - probably even during a single session. The gesture modality

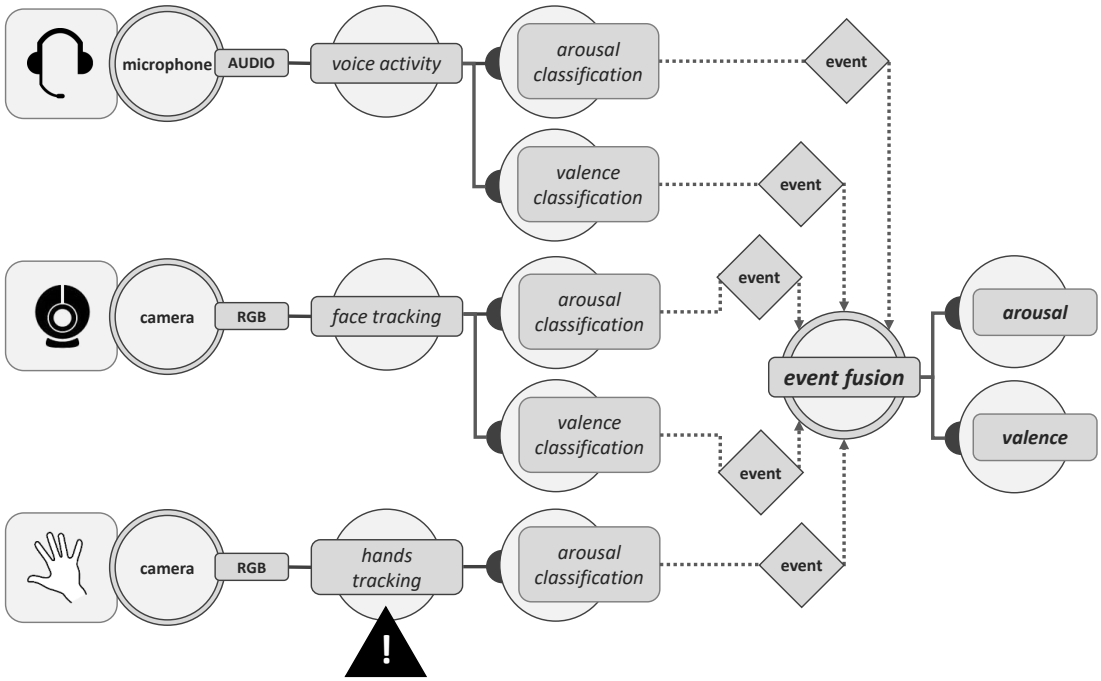


Figure 10.11: The KRISTINA affect recognition system delivers emotional assessments within the valence-arousal space. A special challenge is given by the changing conditions between free-handed and hand-held scenarios, which can be solved by an event-driven fusion approach.

can only be accessed in a free-handed scenario (and even here, the access to gestural information is not guaranteed) but not in the hand-held use case. To our benefit, we have seen in Chapter 9 that an event-driven fusion approach is able to compensate the temporal absence of events from affective channels. The same fusion system can be applied in both use cases, whilst profiting from additional gestural input whenever it becomes available.

Chapter 11

Contributions and Conclusions

The thesis at hand started with the assumption that reasonable combination of information from multiple modalities would have a positive impact on the quality of affect recognition systems. However, the inclusion of several affective channels also offers challenges that have to be addressed by the fusion system in order to generate a positive multi-modal effect.

11.1 Contributions

Considerations how to tackle these problems confronted us with research questions, which we answered within this thesis:

1. The first consideration affects the architecture of a fusion system. A topic often discussed in literature is the difference between early and late fusion. In several experiments we have shown that while applying generalized fusion rules (which can theoretically be used for any given recognition task) there is a certain degree of interchangeability among common fusion strategies, whether they are applied early on the feature level or late on the decision level. In most cases we were only able to achieve a positive multi-modal effect, when we exploited our knowledge about the underlying affect recognition problem and implemented custom, emotion-adapted fusion systems. We can therefore not advise developers to simply apply broadly used fusion rules to their specific affect recognition task, but instead explore the given recognition problem and adapt their fusion schemes accordingly.

2. Going forward from the idea of tailoring the fusion system to the given affect recognition problem, we have evaluated possible ways to model temporally shifted occurrences of affective expressions throughout considered modalities. The classical fusion approach is synchronous, which means that information within a common, equally clocked time interval is considered at each fusion step. We have seen that this naive approach bears the risk of loosing or misinterpreting information that appears in an asynchronous manner across affective channels. In order to deal with this problem, modern fusion systems rely on classification models that implement memory capabilities (e.g. recurrent neural networks). However, these approaches come with the need for immense amount of training data and computational power. As a convenient alternative, we introduced and evaluated the concept of an event driven fusion approach. Here, the fusion component serves as a client which registers affective events that are recognized by uni-modal classifiers. The events are processed with regard to their temporal occurrence. This way temporal dependencies between modalities can be realised and accordingly modelled with event driven fusion algorithms presented in this thesis.

3. The implementation of an event driven fusion approach offers a convenient way to introduce an additional abstraction layer to the affect recognition process. This may be beneficial whenever the given classification task is concerned with complex emotions that are hard to recognize solely based on multi-modal features but can rather be defined by aggregated occurrences or a series of emotional cues. These asynchronous, affect indicating events can be related back to the target emotion by event driven fusion techniques. To elaborate on this hypothesis, we conducted a case study on audiovisual enjoyment recognition and compared event driven approaches relying on uni-modal affective events such as visual smiles and and audible laughs to recurrent neural networks that recognize enjoyment episodes directly from given features. Results endorse the idea that an additional abstraction layer of affective events can indeed improve affect recognition accuracy in comparison to state of the art fusion systems.

All examinations made within this thesis aim at the development of multi-modal affect recognition systems that are suitable for practical application in real-life scenarios. To this end, all presented approaches have been included in the Social Signal Interpretation framework, which supports (near) real-time processing of multi-sensor data as well as machine learning tasks. We have given concrete examples how to handle practical fusion problems especially with an event driven approach with the help of this highly flexible framework. Its mobile version enables the realisation of social signal processing and

affect recognition systems on mobile devices and we consequently brought the event driven fusion approach to the wild. Here, we evaluated a multi-modal fusion application in a real-life scenario, adapting to the challenges offered by a natural setting.

11.2 Conclusions

Our initial goal was the development of an event driven fusion approach and the comparison to other multi-modal affect recognition systems. So after extensive discussions of possible benefits and drawbacks of fusion systems, why and when should event driven fusion be chosen over alternatives?

First off, the eligible alternatives have their merits under the right circumstances. Conventional, synchronous fusion schemes are an easy way to first test the capabilities of additional modalities within an affect recognition system. In an off-line case where we work with segmented sample-lists, that do not contain the session as a whole, temporal relations can only be investigated within the cut-out samples. These time intervals may be too short and there is consequently no other option than to apply a synchronous fusion approach. Assuming we have fully annotated sessions and want to apply asynchronous fusion strategies, recurrent multi-modal networks are the current go-to solution, most often found in present literature. These come in various proposed architectures and provide very reasonable results. However, these neural network approaches require large amounts of training data - unfortunately this precondition is often not provided for affect recognition scenarios (especially if affective channels apart from audiovisual modalities are considered). Furthermore they are often entitled black boxes because of difficulties to trace back the decision making process.

The event driven concept is more accessible and presented algorithms can be well visualised. The running fusion system provides visual feedback, which in addition to improved transparency bears the benefit to ease the tracking of eventual errors in the system. Event driven fusion systems are based on event detection components and therefore implement a modular design. Therefore it is easy to extend with further modalities, its parameters can be hand tuned even after single components are trained and we have seen that it leads to robust recognition systems that can reliably be applied in naturalistic and mobile scenarios. We have shown that event fusion strategies handle tracking and classification errors (which unfortunately can to a certain degree be expected in affect recognition scenarios) well. From an algorithmic point of view, event driven fusion allows asynchronous treatment of modalities, as it is able to include current and

past events from several affective channels in the fusion process. This advantage is shared with state-of-the art fusion systems such as recurrent multi-modal networks. An event driven approach also enables the introduction of an additional abstraction layer in the form of affective events. These target class inducing emotional cues have shown to be able to simplify the classification task when dealing with an abstract target class.

11.3 Future Work

The field of machine learning is a rapidly increasing area of research. The amount of available data is ever growing. Deep and recurrent neural networks especially benefit from this increasing availability of training corpora and the general interest in these algorithms accelerates the development of new ideas how to further exploit these algorithms. We have already seen and compared the application of multi-modal neural networks to the event driven fusion approach, but these strategies do not mutually exclude each other. Actually the exact opposite is the case: Every advance in automatic affect recognition can be taken over one to one for event recognition tasks and in fact are deep neural networks in use for event recognition within the presented research projects. An additional opportunity lies in the recently emerging ways to process the data gained from affective channels with neural networks. Automatic feature learning can be accomplished by feeding raw data into a suitable network topology that is trained towards a target class. The content of layer nodes below the output layer can hereby be interpreted as features - the trained network can consequently be used as feature extraction component. An adjusted approach, recognizing the time intervals in which these features show significant changes with respect to the surrounding data segments could be used to automatically learn affective events from raw data input.

With a growing recognition quality comes the opportunity to refine the event detection process in a way to further subdivide an affective event into a starting, running and ending phase. This can become an important feature to model more complex behaviour patterns. In one of our case examples for event driven fusion, we have seen how an emotion like enjoyment can be recognised by an accumulated occurrence of affective indicator events. Another example showed how the feeling of embarrassment can be described by a very determined sequence of overlapping events. In this case simple accumulation of events may not be enough, start and ending positions become important. The so far presented event driven fusion algorithms may not be sufficient to describe the complex sequence. Algorithms resembling state machines or Markov chains would be better suited to describe the progression of expected events.

In addition to improving the algorithmic approach with advanced learning and processing techniques, we have the ability to tackle further recognition scenarios with the event driven strategy. We have at some points of the thesis indicated the possibility to process affective events of multiple users (e.g. in a conversational scenario) and this possibility enables the analysis of multi-person scenarios. The most straightforward approach would be to assess a group phenomenon such as the enjoyment level of a conversation instead of the assessment of a single user. By having single- and multi-user recognition systems installed in parallel, one could compare the global to the individual observations, e.g. how much each single participant contributes to the enjoyability of the conversation. Interpersonal behaviour where mutual actions or action - reaction patterns describe the sought phenomenon can be well modelled on an event level.

By looking at these few examples, we see that event driven fusion approaches bear good possibilities to profit from developments in the affect recognition field and that multimodal fusion in general will always offer a sophisticated interpretation level between multiple classification results and the final deductions that need to be drawn from these diverse factors. These features will ultimately keep this field of research interesting and important in a future that seems to provide a vast amount of new sensory equipment and available information sources to process.

Bibliography

[1] T. Almaev and M. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013, Geneva, Switzerland, September 2-5, 2013*, 2013, pp. 356–361.

[2] E. Alpaydin and S. Linke, "Maschinelles lernen," 2008.

[3] A. Atkinson, W. Dittrich, A. Gemmell, and A. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," *Perception*, vol. 33, no. 6, pp. 717–746, 2004.

[4] A. Atkinson, M. Tunstall, and W. Dittrich, "Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures," *Cognition*, vol. 104, no. 1, pp. 59–72, 2007.

[5] T. Baltrušaitis, P. Robinson, and L. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, 2016, pp. 1–10.

[6] A. Barreto, J. Zhai, and M. Adjouadi, "Non-intrusive physiological monitoring for automated stress detection in human-computer interaction," in *Human-Computer Interaction, IEEE International Workshop, HCI 2007, Rio de Janeiro, Brazil, October 20, 2007, Proceedings*, 2007, pp. 29–38.

[7] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "Combining efforts for improving automatic classification of emotional user states," *Proc. IS-LTC*, pp. 240–245, 2006.

[8] A. Battocchi, F. Pianesi, and D. Goren-Bar, "Dafex: Database of facial expressions," in *Intelligent Technologies for Interactive Entertainment, First Interna-*

tional Conference, INTETAIN 2005, Madonna di Campiglio, Italy, November 30 - December 2, 2005, Proceedings, 2005, pp. 303–306.

[9] T. Baur, D. Schiller, and E. André, "Modeling users social attitude in a conversational system," in *Emotions and Personality in Personalized Services*. Springer, 2016, pp. 181–199.

[10] N. Bee, E. André, T. Vogt, and P. Gebhard, "The use of affective and attentive cues in an empathic computer-based companion," in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. John Benjamins Publishing Company, 2010, pp. 131–142.

[11] I. Ben-Gal, "Bayesian networks," in *Encyclopedia of statistics in quality and reliability*. Wiley Online Library, 2007.

[12] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martínez, "Emotionet challenge: Recognition of facial expressions of emotion in the wild," *CoRR*, 2017.

[13] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, and R. Pascanu, "Theano: a cpu and gpu math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

[14] T. Bocklet, A. K. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*, 2008, pp. 1605–1608.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[16] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," Computer program (version 5.1.41), July 2010.

[17] F. Boiten, N. Frijda, and C. Wientjes, "Emotions and respiratory patterns: Review and critical analysis," *International Journal of Psychophysiology*, vol. 17, no. 2, pp. 103–128, 1994.

[18] W. Bosma and E. André, "Exploiting emotions to disambiguate dialogue acts," in *Proceedings of the 9th International Conference on Intelligent User Interfaces, IUI 2004, Funchal, Madeira, Portugal, January 13-16, 2004*, 2004, pp. 85–92.

[19] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *1997 Conference on Computer Vision and Pattern Recognition (CVPR'97), June 17-19, 1997, San Juan, Puerto Rico*, 1997, pp. 994–999.

[20] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[21] R. Brueckner and B. Schuller, "Social signal classification using deep blstm recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 4823–4827.

[22] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *International Conference on Multimodal Interfaces (ICMI)*, New York, NY, USA, 2004, pp. 205–211.

[23] S. Card, A. Newell, and T. Moran, *The Psychology of Human-Computer Interaction*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1983.

[24] G. Caridakis, A. Raouzaiou, K. Karpouzis, and S. Kollias, "Synthesizing gesture expressivity based on real sequences," in *LREC 2006 Conference*, 2006.

[25] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in *International Conference on Multimodal Interfaces (ICMI)*, New York, NY, USA, 2006, pp. 146–154.

[26] G. Caridakis, J. Wagner, A. Raouzaiou, Z. Curto, E. André, and K. Karpouzis, "A multimodal corpus for gesture expressivity analysis," in *International Conference on Language Resources and Evaluation (LREC), Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, Malta, 2010.

[27] J. Carroll, "Human-computer interaction: Psychology as a science of design," *Annual review of psychology*, vol. 48, no. 1, pp. 61–83, 1997.

[28] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech." in *Affect and Emotion in Human-Computer Interaction*, ser. Lecture Notes in Computer Science, C. Peter and R. Beale, Eds. Springer, 2008, vol. 4868, pp. 92–103.

[29] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[30] L. Chen, T. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *International Conference on Automatic Face and Gesture Recognition (FGR)*, 1998, pp. 366–371.

[31] G. Chetty, M. Wagner, and R. Goecke, "A multilevel fusion approach for audiovisual emotion recognition." in *AVSP*, 2008, pp. 115–120.

[32] M. Clynes and Y. Menuhin, *Sentics: The touch of emotions*. Anchor Press New York, 1977.

[33] N. Dael, M. Mortillaro, and K. Scherer, "Emotion expression in body action and posture." *Emotion*, vol. 12, no. 5, p. 1085, 2012.

[34] N. Dael, M. Mortillaro, and K. Scherer, "The body action and posture coding system (BAP): Development and reliability," *Nonverbal Behavior*, vol. 36, no. 2, pp. 97–121, 2012.

[35] P. Dagum, A. Galper, and E. Horvitz, "Dynamic network models for forecasting," in *UAI '92: Proceedings of the Eighth Annual Conference on Uncertainty in Artificial Intelligence, Stanford University, Stanford, CA, USA, July 17-19, 1992*, 1992, pp. 41–48.

[36] A. R. Damasio, *Looking for Spinoza: Joy, sorrow, and the feeling brain*. Harcourt, 2003.

[37] B. Dasarathy and B. Sheela, "Composite classifier system design: Concepts and methodology," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 708–713, 1979.

[38] L. De Silva and P. Ng, "Bimodal emotion recognition," in *International Conference on Automatic Face and Gesture Recognition (FGR)*, 2000, pp. 332–335.

[39] C. Demiroglu, D. Anderson, and M. Clements, "A missing data-based feature fusion strategy for noise-robust automatic speech recognition using noisy sensors." in *International Symposium on Circuits and Systems (ISCAS)*, 2007, pp. 965–968.

[40] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, 2011, pp. 2106–2112.

[41] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, vol. 19, no. 3, pp. 34–41, 2012.

[42] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *2013 International Conference on Multimodal Interaction, ICMI '13, Sydney, NSW, Australia, December 9-13, 2013*, 2013, pp. 509–516.

[43] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, November 12-16, 2014*, 2014, pp. 461–466.

[44] A. Dhall, R. Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, 2015, pp. 423–426.

[45] A. Dhall, R. Goecke, J. Joshi, and T. Gedeon, "Emotion recognition in the wild challenge 2016," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, 2016, pp. 587–588.

[46] U. Dimberg, M. Thunberg, and K. Elmehed, "Unconscious facial reactions to emotional facial expressions," *Psychological science*, vol. 11, no. 1, pp. 86–89, 2000.

[47] S. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 147–187, 2010.

[48] S. D'Mello and J. Kory, "Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *International Conference on Multimodal Interaction (ICMI)*, New York, NY, USA, 2012, pp. 31–38.

[49] S. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, p. 43, 2015.

[50] S. Du, Y. Tao, and A. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[51] R. Duin and D. Tax, "Experiments with classifier combining rules," in *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, 2000, pp. 16–29.

[52] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[53] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. Chapman & Hall, 1993.

[54] P. Ekman, *The Philosophy of Deception: Lie Catching and Micro Expressions*. Ed. Clancy Martin, Oxford University Press, 2009.

[55] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1978.

[56] P. Ekman, E. Rosenberg, and J. Hager, "Facial action coding system affect interpretation dictionary (FACSAID)," 1998.

[57] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3, pp. 169–200, 1992.

[58] P. Ekman and W. Friesen, "Constants across cultures in the face and emotion," *Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[59] F. Eyben and B. Schuller, "OpenSMILE:): The Munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

[60] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, 2010, pp. 1459–1462.

[61] F. Eyben, M. Wöllmer, M. Valstar, H. Gunes, B. Schuller, and M. Pantic, "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect," in *International Conference on Automatic Face and Gesture Recognition (FGR)*, USA, March 2011, pp. 322–329.

[62] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[63] S. Flutura, J. Wagner, F. Lingenfelser, A. Seiderer, and E. André, "MobileSSI: Asynchronous fusion for social signal interpretation in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 266–273.

[64] N. Fragopanagos and J. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[65] A. Fridlund, "The new ethology of human facial expressions," *The psychology of facial expression*, vol. 103, 1997.

[66] A. Fridlund, G. Schwartz, and S. Fowler, "Pattern recognition of self-reported emotional state from multiple-site facial emg activity during affective imagery," *Psychophysiology*, vol. 21, no. 6, pp. 622–637, 1984.

[67] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[68] G. Fumera and F. Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 942–956, 2005.

[69] P. Gebhard, "ALMA: a layered model of affect," in *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, New York, NY, USA, 2005, pp. 29–36.

[70] S. Gilroy, M. Cavazza, R. Chaignon, S.-M. Mäkelä, M. Niranen, E. André, T. Vogt, J. Urbain, M. Billinghurst, H. Seichter, and M. Benayoun, "E-tree: Emotionally driven augmented reality art," in *International Conference on Multimedia (MM)*, New York, NY, USA, 2008, pp. 945–948.

[71] S. Gilroy, M. Cavazza, M. Niranen, E. André, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, and M. Billinghurst, "PAD-based multimodal affective fusion," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2009.

[72] M. Glodek, M. Schels, G. Palm, and F. Schwenker, "Multi-modal fusion based on classifiers using reject options and markov fusion networks," in *International Conference on Pattern Recognition (ICPR)*, Nov 2012, pp. 1084–1087.

[73] M. Glodek, M. Schels, F. Schwenker, and G. Palm, "Combination of sequential class distributions from multiple channels using markov fusion networks," *Journal on Multimodal User Interfaces*, vol. 8, no. 3, pp. 257–272, 2014.

[74] M. Glodek, F. Honold, T. Geier, G. Krell, F. Nothdurft, S. Reuter, F. Schüssel, T. Hörnle, K. Dietmayer, W. Minker, S. Biundo, M. Weber, G. Palm, and F. Schwenker, "Fusion paradigms in cognitive technical systems for human-computer interaction," *Neurocomputing*, vol. 161, pp. 17 – 37, 2015.

[75] J. Gratch and S. Marsella, "A domain-independent framework for modeling emotion," *Journal of Cognitive Systems Research*, vol. 5, no. 4, pp. 296–306, 2004.

[76] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[77] D. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[78] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann, 2011.

[79] J. Harrigan, R. Rosenthal, and K. Scherer, *New handbook of methods in nonverbal behavior research*. Oxford University Press, 2008.

[80] B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud, "Design and evaluation of expressive gesture synthesis for embodied conversational agents," in *4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), July 25-29, 2005, Utrecht, The Netherlands*, 2005, pp. 1095–1096.

[81] B. Hartmann, M. Mancini, and C. Pelachaud, "Implementing expressive gesture synthesis for embodied conversational agents," in *Gesture in Human-Computer Interaction and Simulation*, ser. Lecture Notes in Computer Science, S. Gibet, N. Courty, and J.-F. Kamp, Eds. Springer Berlin Heidelberg, 2006, vol. 3881, pp. 188–199.

[82] U. Hess and S. Blairy, "Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy," *International journal of psychophysiology*, vol. 40, no. 2, pp. 129–141, 2001.

[83] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[84] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A field guide to dynamical recurrent neural networks.* IEEE Press, 2001.

[85] J. Hofmann, F. Stoffel, A. Weber, and T. Platt, "The 16 enjoyable emotions induction task (16-EEIT)," 2012, unpublished.

[86] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Technical Report, 2003.

[87] L. Huang, L. Xin, L. Zhao, and J. Tao, "Combining audio and video by dominance in bimodal emotion recognition," in *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, 2007, pp. 729–730.

[88] Y. Huang and C. Suen, "The behavior-knowledge space method for combination of multiple classifiers," in *Conference on Computer Vision and Pattern Recognition, CVPR 1993, 15-17 June, 1993, New York, NY, USA*, 1993, pp. 347–352.

[89] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, 1983, pp. 93–96.

[90] A. Jameel, A. Ghafoor, and M. Riaz, "Improved guided image fusion for magnetic resonance and computed tomography imaging," *The Scientific World Journal*, vol. 13, no. 7, pp. 1–7, 2014.

[91] P. Juslin and K. Scherer, "Vocal expression of affect," in *The new handbook of methods in nonverbal behavior research*. Oxford University Press, 2008, pp. 65–135.

[92] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion space concept," in *2008 16th European Signal Processing Conference, EUSIPCO 2008, Lausanne, Switzerland, August 25-29, 2008*, 2008, pp. 1–5.

[93] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame." *Journal of personality and social psychology*, vol. 68, no. 3, p. 441, 1995.

[94] D. Keltner, P. Ekman, G. Gonzaga, and J. Beer, *Facial expression of emotion.* Oxford University Press, 2003.

[95] J. Kim and E. André, "Emotion-specific dichotomous classification and feature-level fusion of multichannel biosignals for automatic emotion recognition," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI 2008, Seoul, South Korea, August 20-22, 2008*, 2008, pp. 114–119.

[96] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, 2008.

[97] J. Kim and F. Lingenfelser, "Ensemble approaches to parametric decision fusion for bimodal emotion recognition," in *BIOSIGNALS 2010 - Proceedings of the Third International Conference on Bio-inspired Systems and Signal Processing, Valencia, Spain, January 20-23, 2010*, 2010, pp. 460–463.

[98] J. Kim, E. André, M. Rehm, T. Vogt, and J. Wagner, "Integrating information from speech and physiological signals to achieve emotional sensitivity," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, 2005, pp. 809–812.

[99] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single trial classification of eeg and peripheral physiological signals for recognition of emotions induced by music videos," in *International Conference on Brain Informatics*. Springer, 2010, pp. 89–100.

[100] 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*, vol. 3, no. 1, pp. 18–31, 2012.

[101] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, 1995, pp. 1137–1145.

[102] K. Krishna Kishore and P. Krishna Satish, "Emotion recognition in speech using MFCC and wavelet features," in *International Conference on Advance Computing Conference (IACC)*, February 2013, pp. 842–847.

[103] C. Küblbeck and A. Ernst, "Face detection and tracking in video sequences using the modifiedcensus transformation," *Image Vision Comput.*, vol. 24, no. 6, pp. 564–572, 2006.

[104] L. Kuncheva, "Switching between selection and fusion in combining classifiers: An experiment," *Trans. Sys. Man Cyber. Part B*, vol. 32, no. 2, pp. 146–156, April 2002.

[105] L. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, February 2002.

[106] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, "Emotion detection using MFCC and cepstrum features," *Procedia Computer Science*, vol. 70, pp. 29–35, 2015.

[107] P. Lang, M. Bradley, and B. Cuthbert, "Motivated attention: Affect, activation, and action," in *Attention and orienting: Sensory and motivational processes*, P. Lang, R. Simons, and M. Balaban, Eds. Psychology Press, 1997, pp. 97–135.

[108] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011, special issue: Sensing Emotion and Affect - Facing Realism in Speech Processing.

[109] F. Lingenfelser, J. Wagner, T. Vogt, J. Kim, and E. André, "Age and gender classification from speech using decision level fusion and ensemble based techniques," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 2798–2801.

[110] F. Lingenfelser, J. Wagner, and E. André, "A systematic discussion of fusion techniques for multi-modal affect recognition tasks," in *Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, 2011*, 2011, pp. 19–26.

[111] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran, "An event driven fusion approach for enjoyment recognition in real-time," in *Proceedings of the ACM International Conference on Multimedia, MM'14, Orlando, FL, USA, November 03 - 07, 2014*, 2014, pp. 377–386.

[112] F. Lingenfelser, J. Wagner, J. Deng, R. Bruckner, B. Schuller, and E. André, "Asynchronous and event-based fusion systems for affect recognition on naturalistic data in comparison to conventional approaches," *IEEE Transactions on Affective Computing*, 2016.

[113] M. Mancini, G. Varni, D. Glowinski, and G. Volpe, "Computing and evaluating the body laughter index," in *Human Behavior Understanding*. Springer, 2012, vol. 7559, pp. 90–98.

[114] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools and Applications*, vol. 49, no. 2, pp. 277–297, 2010.

[115] H. McCurdy, "Consciousness and the galvanometer." *Psychological Review*, vol. 57, no. 6, p. 322, 1950.

[116] G. McKeown, W. Curran, C. McLoughlin, H. Griffin, and N. Bianchi-Berthouze, "Laughter induction techniques suitable for generating motion capture data of laughter associated body movements," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, Shanghai, China, 22-26 April, 2013*, 2013, pp. 1–5.

[117] G. McKeown, W. Curran, J. Wagner, F. Lingenfelser, and E. André, "The Belfast storytelling database: A spontaneous social interaction database with laughter focused annotation," in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*, 2015, pp. 166–172.

[118] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states." *Genetic, social, and general psychology monographs*, vol. 121, no. 3, pp. 339–361, 1995.

[119] A. Mehrabian and S. Ferris, "Inference of attitudes from nonverbal communication in two channels." *Consulting Psychology*, vol. 31, no. 3, pp. 248–252, 1967.

[120] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[121] A. Mehrabian and M. Wiener, "Decoding of inconsistent communications." *Journal of personality and social psychology*, vol. 6, no. 1, p. 109, 1967.

[122] M. Mortillaro, B. Meuleman, and K. Scherer, "Advocating a componential appraisal model to guide emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 3, no. 1, pp. 18–32, 2012.

[123] C. Müller, "Automatic recognition of speakers age and gender on the basis of empirical studies," in *Proceedings of the 9th International Conference on Spoken Language Processing. Conference in the Annual Series of Interspeech Events (INTERSPEECH-06), September 17-21, Pittsburg,, PA, USA*, 2006.

[124] F. Nasoz, K. Alvarez, C. Lisetti, and N. Finkelstein, "Emotion recognition from physiological signals using wireless sensors for presence technologies," *Cognition, Technology & Work*, vol. 6, no. 1, pp. 4–14, 2004.

[125] A. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2002, May 13-17 2002, Orlando, Florida, USA*, 2002, pp. 2013–2016.

[126] M. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, 2010, pp. 3695–3699.

[127] R. Niewiadomski, M. Mancini, T. Baur, G. Varni, H. Griffin, and M. Aung, "Mmli: Multimodal multiperson corpus of laughter in interaction," in *Human Behavior Understanding*, ser. Lecture Notes in Computer Science, A. Salah, H. Hung, O. Aran, and H. Gunes, Eds. Springer International Publishing, 2013, vol. 8212, pp. 184–195.

[128] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrusaitis, and L. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, 2016, pp. 284–288.

[129] P. Omedas, A. Betella, R. Zucca, X. D. Arsiwalla, D. Pacheco, J. Wagner, F. Lingenfelser, E. Andre, D. Mazzei, A. Lanatá *et al.*, "Xim-engine: a software framework to support the development of interactive applications that uses conscious and unconscious reactions in immersive mixed reality," in *Proceedings of the 2014 Virtual Reality International Conference, VRIC 2014, Laval, France, April 9-11, 2014*. ACM, 2014, pp. 26:1–26:4.

[130] H. Pan, S. Levinson, T. Huang, and Z.-P. Liang, "A fused hidden markov model with application to bimodal speech processing," *IEEE Transactions on Signal Processing*, vol. 52, no. 3, pp. 573–581, 2004.

[131] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, September 2003.

[132] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang, "Human computing and machine understanding of human behavior: A survey," in *Artifical Intelligence for Human Computing, ICMI 2006 and IJCAI 2007 International Workshops, Banff, Canada, November 3, 2006, Hyderabad, India, January 6, 2007, Revised Seleced and Invited Papers*, 2007, pp. 47–71.

[133] R. Picard, *Affective computing*. Cambridge, MA, USA: MIT Press, 1997.

[134] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001.

[135] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 3, pp. 21–45, 2006.

[136] C. Puri, L. Olson, I. Pavlidis, J. Levine, and J. Starren, "Stresscam: non-contact measurement of users' emotional states through thermal imaging," in *Extended Abstracts Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005, Portland, Oregon, USA, April 2-7, 2005*, 2005, pp. 1725–1728.

[137] D. Quintana, A. Guastella, T. Outhred, I. Hickie, and A. Kemp, "Heart rate variability is associated with emotion recognition: Direct evidence for a relationship between the autonomic nervous system and social cognition," *International Journal of Psychophysiology*, vol. 86, no. 2, pp. 168–172, 2012.

[138] A. Rabie, B. Wrede, T. Vogt, and M. Hanheide, "Evaluation and discussion of multi-modal emotion recognition," in *Proceedings of the 2009 Second International Conference on Computer and Electrical Engineering - Volume 01*, Washington, DC, USA, 2009, pp. 598–602.

[139] S. Raudys and F. Roli, "The behavior knowledge space fusion method: Analysis of generalization error and strategies for performance improvement," in *Multiple Classifier Systems, 4th International Workshop, MCS 2003, Guilford, UK, June 11-13, 2003, Proceedings*, 2003, pp. 55–64.

[140] F. Ringeval, F. Eyben, E. Kroupi, A. Yüce, J. Thiran, T. Ebrahimi, D. Lalanne, and B. W. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[141] F. Rosenblatt, *Principles of Neurodynamics*. Spartan, 1963.

[142] T. Ruf, A. Ernst, and C. Küblbeck, "Face detection with the sophisticated high-speed object recognition engine (SHORE)," in *Microelectronic Systems*. Springer, 2011, pp. 243–252.

[143] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error propagation," DTIC Document, Technical Report, 1985.

[144] B. Schuller, R. Müller, M. K. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, 2005, pp. 805–808.

[145] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 2794–2797.

[146] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 3201–3204.

[147] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, 2012, pp. 254–257.

[148] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in

INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, 2013, pp. 148–152.

[149] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 427–431.

[150] N. Sebe, I. Cohen, T. Gevers, and T. Huang, "Multimodal approaches for emotion recognition: A survey," in *Internet Imaging VI*, December 2004, pp. 56–67.

[151] A. Seewald and J. Fürnkranz, "An evaluation of grading classifiers," in *Advances in Intelligent Data Analysis, 4th International Conference, IDA 2001, Cascais, Portugal, September 13-15, 2001, Proceedings*, 2001, pp. 115–124.

[152] G. Shafer, *A Mathematical Theory of Evidence*. Princeton: NJ: Princeton Univ. Press, 1976.

[153] B. Shneiderman, *Software psychology: Human factors in computer and information systems (Winthrop computer systems series)*. Winthrop Publishers, 1980.

[154] B. Shneiderman, "Putting the human factor into systems development," in *Proceedings of the eighteenth annual computer personnel research conference*. ACM, 1981, pp. 1–13.

[155] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[156] M. Soleymani, "Detecting cognitive appraisals from facial expressions for interest recognition," *CoRR*, vol. abs/1609.09761, 2016.

[157] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological signals (TEAP)," *Front. ICT*, 2017.

[158] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition - A new approach," in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA*, 2004, pp. 1020–1025.

[159] K. Ting and I. Witten, "Issues in stacked generalization," *J. Artif. Intell. Res.*, vol. 10, pp. 271–289, 1999.

[160] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 5200–5204.

[161] M. Valderas, J. Bolea, P. Laguna, M. Vallverdú, and R. Bailón, "Human emotion recognition using heart rate variability analysis with spectral bands based on respiration," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015, Milan, Italy, August 25-29, 2015*, 2015, pp. 6134–6137.

[162] E. Velten, "A laboratory task for induction of mood states," *Behavior Research and Therapy*, vol. 6, pp. 473–482, 1968.

[163] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Trans. Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[164] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, The Netherlands*, 2005, pp. 474–477.

[165] T. Vogt and E. André, "Exploring the benefits of discretization of acoustic features for speech emotion recognition," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 328–331.

[166] T. Vogt, E. André, and N. Bee, "Emovoice - A framework for online recognition of emotions from voice," in *Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008, Kloster Irsee, Germany, June 16-18, 2008, Proceedings*, 2008, pp. 188–199.

[167] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification,"

in *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, The Netherlands*, 2005, pp. 940–943.

[168] J. Wagner, F. Lingenfelser, E. André, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affective Computing*, vol. 2, no. 4, pp. 206–218, 2011.

[169] J. Wagner, F. Lingenfelser, and E. André, "Using phonetic patterns for detecting social cues in natural conversations," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 168–172.

[170] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time," in *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, 2013, pp. 831–834.

[171] J. Wagner, F. Lingenfelser, D. Mazzei, A. Betella, R. Zucca, P. Omedas, and P. Verschure, "A sensing architecture for empathetic data systems," in *4th Augmented Human International Conference, AH'13, Stuttgart, Germany, March 7-8, 2013*, 2013, pp. 96–99.

[172] J. Wagner, F. Lingenfelser, and E. André, *Building a Robust System for Multimodal Emotion Recognition*. Wiley, 2015, ch. 15, pp. 379–410.

[173] L. Wanner, E. André, J. Blat, S. Dasiopoulou, M. Farrùs, T. Fraga, E. Kamateri, F. Lingenfelser, G. Llorach, O. Martínez *et al.*, "KRISTINA: A knowledge-based virtual conversation agent," in *Advances in Practical Applications of Cyber-Physical Multi-Agent Systems: The PAAMS Collection - 15th International Conference, PAAMS 2017, Porto, Portugal, June 21-23, 2017, Proceedings*, 2017, pp. 284–295.

[174] L. Wanner, E. André, J. Blat, S. Dasiopoulou, M. Farrús, T. Fraga, E. Kamateri, F. Lingenfelser, G. Llorach, O. Martínez *et al.*, "Design of a knowledge-based agent as a social companion," *Procedia Computer Science*, vol. 121, pp. 920–926, 2017.

[175] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT: The Munich open-source CUDA recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015.

[176] P. Weyers, A. Mühlberger, C. Hefele, and P. Pauli, "Electromyographic responses to static and dynamic avatar emotional facial expressions," *Psychophysiology*, vol. 43, no. 5, pp. 450–453, 2006.

[177] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[178] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Neurocomputing*, vol. 73, no. 1-3, pp. 366–380, 2009.

[179] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening." *Selected Topics Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

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

[181] D. Wolpert and W. Macready, "Coevolutionary free lunches," *IEEE Trans. Evolutionary Computation*, vol. 9, no. 6, pp. 721–735, 2005.

[182] L. Wu, S. Oviatt, and P. Cohen, "Multimodal integration - A statistical view," *IEEE Trans. Multimedia*, vol. 1, no. 4, pp. 334–341, 1999.

[183] Z. Xue and R. Blum, "Concealed weapon detection using color image fusion," in *Proceedings of the 6th International Conference on Information Fusion*, 2003, pp. 622–627.

[184] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T. Huang, D. Roth, and S. Levinson, "Bimodal HCI-related affect recognition," in *Proceedings of the 6th International Conference on Multimodal Interfaces, ICMI 2004, State College, PA, USA, October 13-15, 2004*, 2004, pp. 137–143.

[185] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, 2005, pp. 967–972.

[186] Z. Zeng, J. Tu, B. Pianfetti, and T. Huang, "Audio-visual affective expression recognition through multistream fused HMM," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.

[187] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[188] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 1, pp. 115–126, 2015.