Asynchronous and Event-Based Fusion Systems for Affect Recognition on Naturalistic Data in Comparison to Conventional Approaches

Florian Lingenfelser, Johannes Wagner, Jun Deng, Raymond Brueckner, Björn Schuller, *Member, IEEE*, and Elisabeth André

Abstract—Throughout many present studies dealing with multi-modal fusion, decisions are synchronously forced for fixed time segments across all modalities. Varying success is reported, sometimes performance is worse than unimodal classification. Our goal is the synergistic exploitation of multimodality whilst implementing a real-time system for affect recognition in a naturalistic setting. Therefore we present a categorization of possible fusion strategies for affect recognition on continuous time frames of complete recording sessions and we evaluate multiple implementations from resulting categories. These involve conventional fusion strategies as well as novel approaches that incorporate the asynchronous nature of observed modalities. Some of the latter algorithms consider temporal alignments between modalities and observed frames by applying asynchronous neural networks that use memory blocks to model temporal dependencies. Others use an indirect approach that introduces events as an intermediate layer to accumulate evidence for the target class through all modalities. Recognition results gained on a naturalistic conversational corpus show a drop in recognition accuracy when moving from unimodal classification to synchronous multimodal fusion. However, with our proposed asynchronous and event-based fusion techniques we are able to raise the recognition system's accuracy by 7.83 percent compared to video analysis and 13.71 percent in comparison to common fusion strategies.

Index Terms—Affective computing, multimodal recognition, sensor fusion, artificial neural networks, deep learning

1 Introduction

In unimodal 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 [9]. But since the cues which describe emotional conditions are indeed encoded within multiple modalities, the classification process should incorporate as much multimodal information as possible from *multiple* channels [43]. Elaborate ways of fusing multiple modalities are in use throughout many affect recognition studies. A number of studies [27], [34], [40], [41], [42] combine differing modalities and mostly confirm the assumption that multimodal fusion leads to more accurate affect recognition systems than a unimodal approach. Literature offers several surveys, which compare multiple approaches to fusion of multimodal information for affect recognition [28], [43]. D'Mello et al. [5] give a very detailed overview over 30 studies which deal with multimodal emotion recognition systems and came up with the following observations. Only one quarter of studies deal with natural corpora and spontaneous emotion, three quarters work on acted data. Used corpora directly influence the success of fusion systems: recognition improvements range from −9 to +27 percent, whereby the positive effect on accuracy is three times higher on acted emotions. Furthermore, the impact of a multimodal approach can be roughly predicted by the performance of the best unimodal classifier. This finding implies that a certain emotion may be visible in one affective channel, but has a modest effect in additional modalities. The application of affect recognition in real-world scenarios requires the used algorithms to perform well on naturalistic input. This means that more subtle cues than the ones shown in acted and partially exaggerated affective states have to be recognized and interpreted. Furthermore, the simultaneous display of an emotion in all affective channels is at least unlikely to happen in a naturalistic setting.

There is also evidence from psychological studies that the temporal dynamics of emotional displays are not necessarily the same across all modalities. To illustrate this, let us have a look at a typical behavioral pattern expressing embarrassment. According to Keltner et al. [17], the display of embarrassment usually starts with a gaze down followed by a sequence of smiles, gaze and head shifts. 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.

- *F. Lingenfelser, J. Wagner and E. André are with the Lab for Human Centered Multimedia, HCM, Universität Augsburg, Augsburg 86159, Germany. E-mail: {lingenfelser, wagner, andre}@hcm-lab.de.*
- *J. Deng, R. Brueckner and B. Schuller are with the Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, München 80333, Germany.*
  E-mail: jundeng86@gmail.com, raymond.brueckner@web.de, schuller@tum.de.

In practice, however, fusion algorithms are often applied in a synchronous way: 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 and during this time the facial expression is also considered. 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. So how do we solve the problem of non-aligned cues in multiple signals in order to recognize natural emotions in a better way?

A first step is to reject the assumption that all relevant cues happen in all modalities at the same time, which is an implicit assumption of the synchronous fusion approaches described so far. We need to apply fusion strategies that address the asynchronous nature of the recognition problem. An obvious way to relate temporally unaligned cues for proper fusion is the use of dynamic classifiers. Song et al. [34] describe a tripled Hidden Markov Model, which is able to integrate three streams of data and allows for the state asynchrony of the sequences while preserving their natural correlation over time. More elaborate ways of applying asynchronous fusion to multimodal data streams can be found in the form of recurrent neural networks with memory capabilities [3], [14]. They are able to learn a history of past frames in several modalities and take them into account for classification. The mentioned fusion mechanisms respect the asynchronicity of multiple modalities; however, they apply classification directly: Descriptive features of observed channels are used to train classifiers to directly recognize the sought target class.

Event-based fusion approaches offer an indirect way to tackle affect recognition from several modalities by recognizing events as indicators for the target class and accumulating their asynchronous occurrences within the affective channels over time. In [21], a vector based approach is presented that calculates the target class probability from asynchronous indicator events, by relating back recognized visual smiles and audible laughs to a positive emotional state, instead of trying to classify it directly from features gained from video and audio analysis. The fusion approach was evaluated by recognizing a speaker's level of enjoyment in a natural conversation scenario. 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 [29]. Enjoyment can be defined as an episode of enjoyable emotion. These episodes are typically accompanied by visual and auditory cues, which makes this classification task a well-suited proving ground for recognition systems that make use of multimodal fusion techniques. In [21], we did only compare a single event-based fusion algorithm to conventional recognition approaches. This evaluation will now be extended by looking at other sophisticated asynchronous fusion strategies in the form of recurrent neural networks as well as the introduction of novel event-based algorithms that e.g., use Dynamic Bayesian Networks for event processing. Therefore we will carry out a systematic categorization of proposed fusion strategies (Section 2) and an evaluation (Section 5), in which we compare proposed asynchronous (Section 3.2) and event-based (Section 3.3) fusion approaches to conventional affect recognition approaches (Section 3.1) for framewise enjoyment recognition on naturalistic data (Section 4).

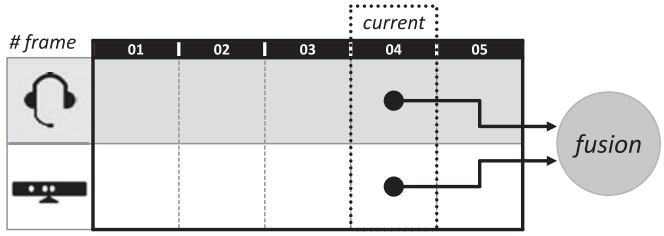


Fig. 1. Synchronous fusion approaches are characterized by the consideration of multiple modalities within the same time frame.

2 Categorization of Fusion Systems

Meta studies that describe and compare several approaches for affect recognition often try to categorize the used fusion algorithms. Shivappa et al. [33] differentiate between fusion approaches at the signal, feature, model, decision and semantic level. These are mostly synchronous fusion approaches and we will cover most of these schematics in Section 3.1. Glodek et al. [12], [11] describe fusion architectures that apply fusion gradually at different levels, including fusion steps from levels of signal recognition to abstract logical inferences. They therefore describe the three categories of perception-level fusion, knowledge-based fusion and application-level fusion. Concerning this classification of fusion levels, the fusion systems discussed in this article lie within the perception-level. In addition to these characteristics, the applied recognition schemes may be distinguished by two major decisions that arise when designing a multimodal affect recognition system.

2.1 Synchronous versus Asynchronous Recognition

Confronted with the task of fusing multiple affective modalities, a vast amount of eligible strategies come into consideration [30]. In this study, we differentiate fusion techniques by the way they handle temporal alignments between information within modalities. Synchronous fusion approaches share the characteristic of considering multiple modalities and respective feature sets within the same time slice (Fig. 1). Zeng et al. [43] cite 18 studies dealing with audio-visual fusion distinguishing between feature-, decision-, and model-level fusion. Feature-level fusion is a very straightforward way to fuse all observed modalities by merging all calculated features into a single and high dimensional feature set to form one single classification model. The multimodal feature set contains a greater amount of information than a single modality. Decision-level fusion sums up combination

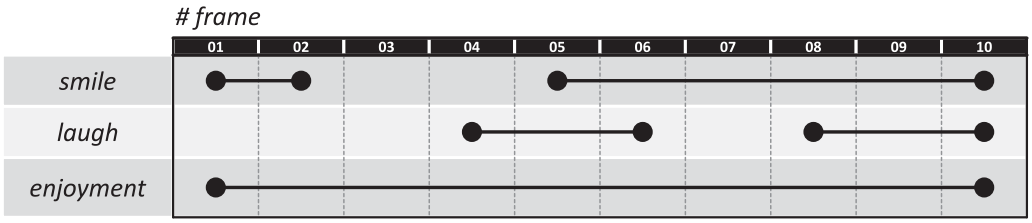


Fig. 2. Exemplary annotations of enjoyment, voiced laughs and visual smiles. Dotted lines depict time frames in which decisions have to be made by the fusion system. Asynchronous and even-based fusion approaches have the opportunity to overcome segments with a sparse distribution of actual cues of enjoyment.

rules for the probabilistic outputs of several classification models. Instead of using all available features for a single classifier, the available feature set is divided into subgroups (e.g., one classifier per modality). Standard decision techniques include class-label combination (e.g., voting or look-up tables) and algebraic combination rules (e.g., sum rule or product rule). Feature- and decision-level fusion include the most standard approaches used in many studies concerning multimodal fusion experiments. In model-level fusion, such as Stacked Generalisation [36], the outputs of several 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 decision. Studies, such as [6], [10], [18], [19], [20], examine rather basic fusion strategies and sometimes advise which scheme dominates others.

Asynchronous fusion is meant to not force decisions on synchronised time frames (Fig. 3). Dupont et al. [7] were among the first to tackle the asynchronous nature of audio and video streams by modelling temporal topologies with multi-stream HMMs for continuous speech recognition. Zeng et al. [42] applied Multi-stream Fused Hidden Markov Model (MFHMM), where state transitions of different component HMMs do not necessarily occur at the same time across different streams so that the synchrony constraint among different streams is also relaxed. Coupled Hidden Markov Models (CHMM) [26] have also been proposed. Here the probability of the next state of a sequence depends on the current state of all HMMs and therefore enables an improved modelling of intrinsic temporal correlations between multiple modalities. To overcome the computational complexity of asynchronous Hidden Markov model (AHMM), Wöllmer et al. [40] suggested a multidimensional dynamic time warping (DTW) algorithm for hybrid fusion of asynchronous data, requiring significantly less decoding time while providing the same data fusion flexibility as the AHMM. Finally, Recurrent Neural Networks (RNNs) offer a third alternative for asynchronous fusion, in particular in the form of Long Short-Term Memory Neural Networks (LSTM-NNs), which replace the traditional neural network nodes with memory cells, essentially allowing the network to learn when to store or relate to bimodal information over long periods of time. We have successfully applied LSTM-NNs to combine acoustic and linguistic features to continuously predict the current quadrant in a two-dimensional emotional space spanned by the dimensions valence and activation [41]. Likewise, in a similar emotion recognition task, this approach successfully fuses facial expressions, shoulder gestures and audio cues [28].

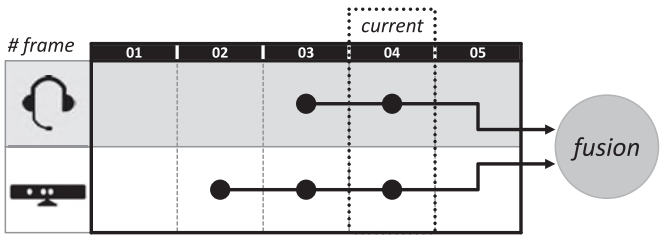


Fig. 3. Asynchronous fusion scheme. Unlike synchronous fusion, which considers multiple modalities within the same (current) time frame, asynchronous fusion algorithms refer to past time frames with the help of some kind of memory support. Therefore they are able to catch temporal shifts of emotional expressions between observed modalities.

2.2 Expected Gains of Asynchronous Approaches

Fig. 2 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 multimodal 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 the modalities into account, should be able to catch the characteristics of the analysed data more precisely. 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.

2.3 Direct versus Event-Based Recognition

Most approaches to emotion recognition are based on categorical emotion theories [8], which model emotions as distinct categories, such as joy, anger, surprise, fear or sadness, or dimensional emotion theories, which characterize emotions in terms of several continuous dimensions [23], such as pleasure, arousal and dominance. Classifiers are usually trained to map relevant features directly on discrete emotion categories or on a continuous multidimensional space. Typically the probabilistic output of modality-specific classifiers is combined by a fusion strategy and an agreeing decision is determined among the considered modalities. These approaches may be referred to as direct emotion recognition approaches.

In this paper, we introduce events as an intermediate layer of representation. Instead of directly recognizing affective states from relevant features, we search for indicative events that can be algorithmically interpreted for target

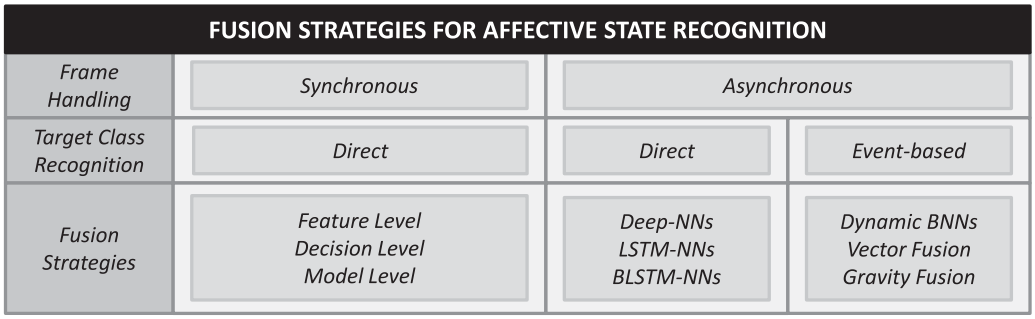


Fig. 4. We can hierarchically group the appropriate fusion strategies (Section 3) depending on the decisions made for the treatment of the temporal dynamics and the levels of processing. The first layer depicts the distinction between a synchronous or asynchronous approach (Section 2.1). Second, classifiers 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 (Section 2.3).

class estimation. In the example above, audio features can be used to classify laugh events and their repeated occurrence can be taken as a hint of enjoyment. Similar events can be defined for all available modalities and found events can consequently be fed into combination algorithms that relate these indicator events back to the target class (Fig. 5).

An intermediate layer of representation has also been suggested by Mortillaro et al. [24] 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 which lie between expressive features and emotional labels.

2.4 Expected Gains from Event-Based Recognition

The promise of the event-based approach lies in the introduction of an intermediate layer of representation that reduces the complexity of the emotion recognition task while enhancing its transparency. For example, Fig. 2 shows how enjoyment is recognized from a repetitive sequence of voice laughs and visual smiles. The figure gives an impression of which events to expect when observing enjoyment. Even though voice laughs and visual smiles are correlated, they do not start and end at the same point in time. Also there are phases of enjoyment where neither a voice laugh nor a visual smile occurs. The benefit of the event-based approach is that it allows for a sufficient amount of flexibility to represent the typical occurrences of voice laughs and visual smiles in enjoyment. In particular, it does not force us to code multimodal emotional behaviors on an exact time line, but instead accumulates evidence from relevant events as they occur in order to recognize the target class.

3 Implemented Fusion Strategies

In the previous section, we have presented a taxonomy of fusion strategies (see Fig. 4) depending on the treatment of timing and the levels of processing. We have discussed the potential benefits of asynchronous over synchronous fusion approaches and argued in favor of an intermediate layer between low-level features and high-level affective states. In this section, concrete implementations of respective algorithms will be presented to investigate the assumptions made in more detail.

3.1 Synchronous Fusion

Feature-, decision-, and model-level fusion schemes clearly are among the most reported strategies for combining multiple modalities for affect recognition [5]. These algorithms are generally easy to implement and apply, which makes them an obvious choice for the synchronous fusion approach. As discussed in Section 2.1, feature-level fusion is by far the most straightforward combination strategy, but nevertheless has often proven to yield good results in comparison to unimodal classification and other fusion schemes [20]. Contrary to feature-level fusion, decision-level fusion focuses on the usage of several classifiers and combination of their probabilistic outputs. As a baseline for the comparison, we choose the widely used product rule for merging the output of several classification models: The decision of ensemble member t for class n is denoted as $d_{t,n} \in \{0,1\}$, with $t = 1 \dots T$ and $n = 1 \dots N$ 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 $s_{t,n} \in [0,1]$. By multiplying the support given to each class ω_n, total support μ_n for class n is calculated as

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

The ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest. Within the comparison study, we also observe the possibility of

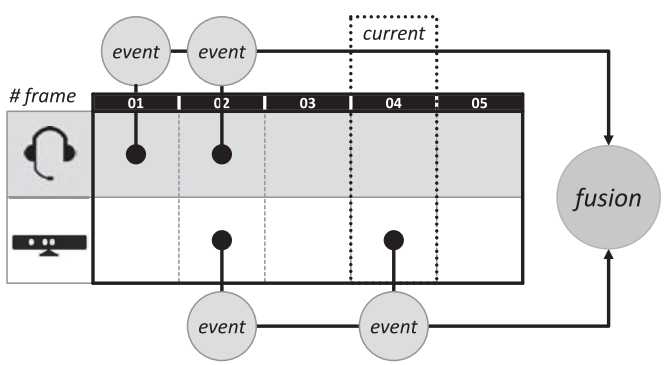


Fig. 5. Event-based fusion scheme. The target class is not directly classified, but target class indicating events are recognized by accordingly trained models. The final classification has to be algorithmically derived from found events.

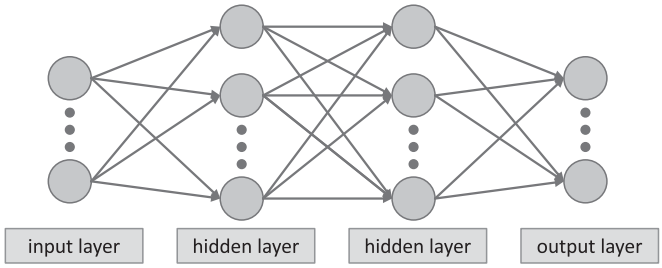


Fig. 6. Multilayer perceptron with input layer, several hidden layers and an output layer containing the classification result.

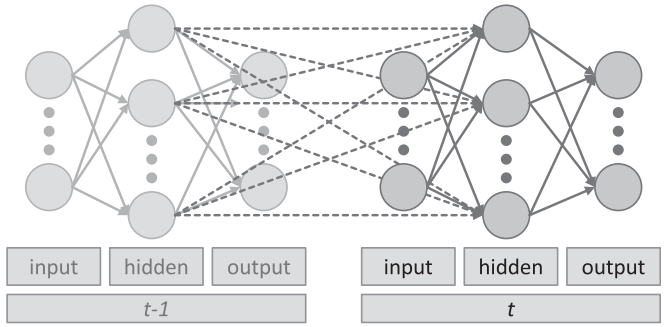


Fig. 7. Recurrent neural network with input layer, one hidden layer and an output layer containing the classification result. The hidden layer at time *t* has self-connections to the remembered hidden layer of *t-1*.

model-level fusion, in which the probabilistic 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. In detail, we apply the technique of Stacked Generalisation as proposed by [36]. One meta classifier tries to learn the probability distribution among ensemble classifiers together with the true class that leads to this combination. When asked to classify an unknown sample x, the method first collects probability estimates of all ensemble members that consecutively form the basis for the meta classifier's final prediction.

3.2 Direct Asynchronous Fusion

Neural networks were initially designed to simulate the human brain's learning processes as machine learning scheme [31]. They describe a network of nodes (neurons), linked by weighted connections (synapses). In a multilayer perceptron [32], multiple hidden layers connect the input layer to the output layer-the classification result (Fig. 6). The modifications explained in this section make neural networks a very potent choice for asynchronous classification, and therefore we apply them as a strategy for direct asynchronous fusion within our comparisons (Fig. 3).

Deep neural networks (Deep-NNs), which contain a large number of parameters, have recently become the gold standard for a various of applications, such as speech recognition and image classification. We therefore compare Deep-NN-based methods with our proposed methods. In particular, deep feedforward neural networks with rectifier neural units are used to build the recognition model in [25].

In order to better capture the asynchronicity of modalities, Deep-NNs can be enhanced with memory capabilities: Recurrent neural networks (RNNs) are characterized by having cyclic connections and receive their input not only from the input layer, but also from hidden nodes that *remember* previous time steps (Fig. 7). Though at this stage, RNNs are capable of handling temporal alignments, the range of temporal information that the recurrent network can access is limited [15]. This problem is caused by the so-called *vanishing gradient problem*, which is the phenomenon of exponentially decaying influence from input to hidden and output layers.

Long Short-Term Memory neural networks (LSTM-NNs) are devised to better handle and exploit temporal context by using memory blocks [14], which consist of several recurrently connected subnets: Each memory block contains one or more recurrently connected memory cells and three gate units, namely the input, output, and forget gates, which control the information flow to and inside the memory block and are designed to solve the *vanishing gradient problem* (Fig. 8). LSTM-NNs based models for asynchronous fusion have shown remarkable success in paralinguistic tasks [3], [41]. For further comparison with our proposed methods, we train deep LSTM models for enjoyment recognition. In addition, bi-directional LSTM-NNs (BLSTM-NNs) are implemented, as these neural networks can access and utilize past and future context and can therefore be expected to yield the best recognition results of neural networks considered in this study.

3.3 Event-Based Fusion

Direct asynchronous fusion approaches theoretically outperform synchronous techniques by modelling temporal relations across modalities. However, they are trained to directly recognize the target class. This means the system considers every time slice and decides if it belongs to the sought class (es), not taking into account whether the observed time frame contains expressive information at all. Event-based fusion takes another approach to classification: Separately trained recognizers look for target class indicating events in the considered modalities and report their occurrence and probability to the fusion algorithm. The fusion system accumulates registered events, considers their temporal alignments and deduces the target class likelihood from the given information. Each modality serves as a client which individually decides when to add information. Signal processing components can be added or replaced without having to touch the actual fusion system and missing input from one

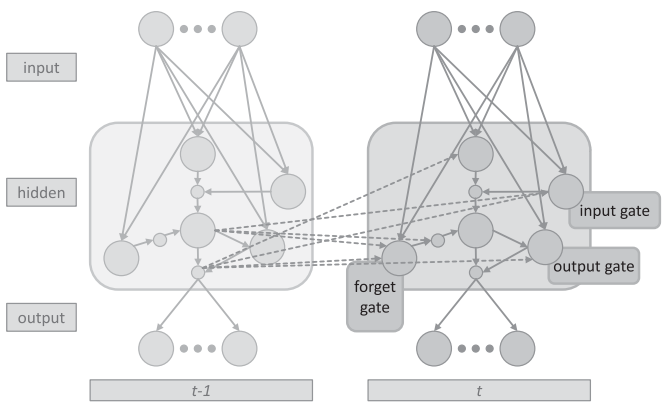


Fig. 8. LSTM network with one memory block, including the input, output, and forget gate.

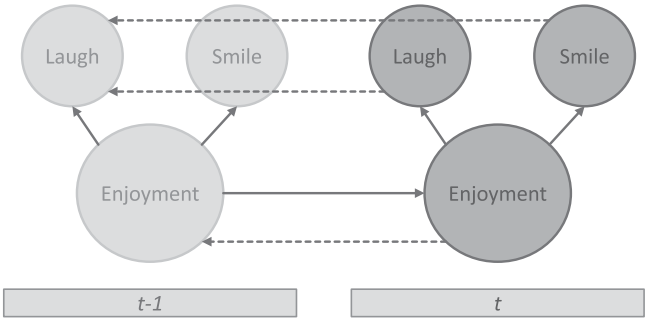


Fig. 9. Structure of a dynamic Bayesian network for enjoyment estimation. Each frame laugh and smile nodes *t* are updated with current events and outdated confidence values are shifted shifted one time slice into the past *t-1* (dotted arrows). Enjoyment estimation is therefore calculated from current observations and probability distributions of past frames.

of the modalities does not cause the collapse of the whole fusion process and adds to a very good expandability of the final affect recognition system.

In Section 3.2 we described the possibilities to model temporal relations between modalities with artificial neural networks. Although proven to be capable of handling asynchronicity, the resulting networks are rather complex black-boxes. Once trained it is hard to adjust them to new conditions when applying them within a naturalistic setting to which the learned model parameters may poorly generalize (although there exist techniques like dropout [35], which help to alleviate problems such as overfitting on training data). Event based fusion approaches like the algorithms described in Sections 3.3.2 and 3.3.3 have a set of comprehensible parameters (e.g., the decay speed of events) that can on the one hand be learned and optimized with training data, but can on the other hand also be tuned by expert knowledge to quickly react to new conditions in a real-time scenario.

3.3.1 Dynamic Bayesian Networks

Fig. 9 shows the structure of a Dynamic Bayesian Network (Dynamic-BN), which is used to collect laugh and smile events, tagged with respective confidence values. From these it calculates enjoyment probability based on current and preceding observations: Every frame, modalities are checked for occurring enjoyment indicating laugh and smile 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 laugh, smile and enjoyment nodes (*t*) is shifted one time slice into the past (*t-1*). Probability of enjoyment within the current frame is subsequently calculated from current observations and probability distributions of past frames. Initial configuration is learnt from framewise annotations of the used corpus (Fig. 2) and models general distributions of frames containing laughs, smiles and enjoyment episodes.

3.3.2 Vector Fusion

The proposed fusion algorithm is based on preceding work done by [13], which represents emotions by means of pleasure and arousal in a two-dimensional emotional space. 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, as a confidence value is defined for each axis in the event space. This defines the position of the vector within the dimensional model. We dynamically calculate it from the normed probabilities of a recognized cue, resulting in values that 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 zero. 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:

$$decay_e = norm_e - (lifetime_e * speed_e).$$

If the resulting norm of the decayed vector stays above zero, it remains active-otherwise the vector is discarded. Afterwards a mass centre is calculated over all active event vectors:

For each dimension $d = 1 \dots D$ of the vector space respective values of active event vectors $e = 1 \dots E$ (modified by their weight factor) are summed up

$$mass(d_{1\dots D}) = \sum_{e=1}^{E} (eventvector_{d,e} * weight_e).$$

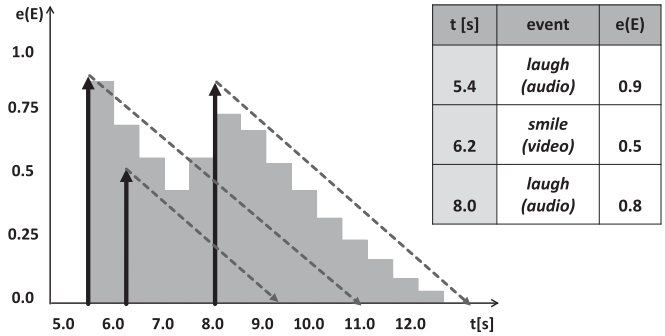
The result is normalized by the sum of weights of all contributing event vectors

$$mass(d_{1\dots D}) = mass(d_{1\dots D}) \Big/ \sum_{e=1}^{E} weight_e.$$

The fusion result itself is a vector which approaches the calculated mass centre with a predefined speed parameter (Fig. 10). If this vector rises above a specified threshold, we classify the frame to contain the target class. If no events remain active in the vector space, the fusion vector approaches a neutral state.

3.3.3 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 an n-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



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

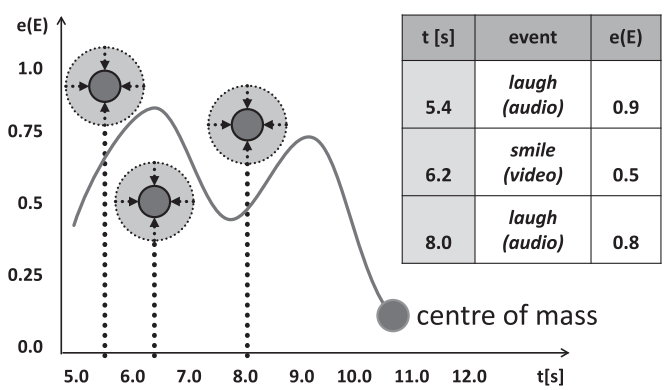
Fig. 10. Example schematic of the vector fusion algorithm. Three enjoyment indicating events from audio and video modality (black arrows) are successively mapped into the vector space. Their lengths decrease over time (dotted lines), therefore the mass centre moves over time with the decreasing vectors.

temporal decay of the weight of mass points (Fig. 11). 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.

4 Naturalistic Multimodal Data

For the comparison study of affect recognition strategies which we will present in the following sections, we use the first session of the Belfast Storytelling Database [22]. 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. This fact makes the corpus well suited for evaluating the implemented fusion approaches (Section 3), because it enables the comparison of direct classification to the recognition and fusion of multimodal indicator events.

The corpus consists of six sessions of groups of three or four people telling stories to one another in either English or Spanish. Each session lasts about 120 minutes, resulting in approximately 75 minutes recording time. The storytelling



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

Fig. 11. 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 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.

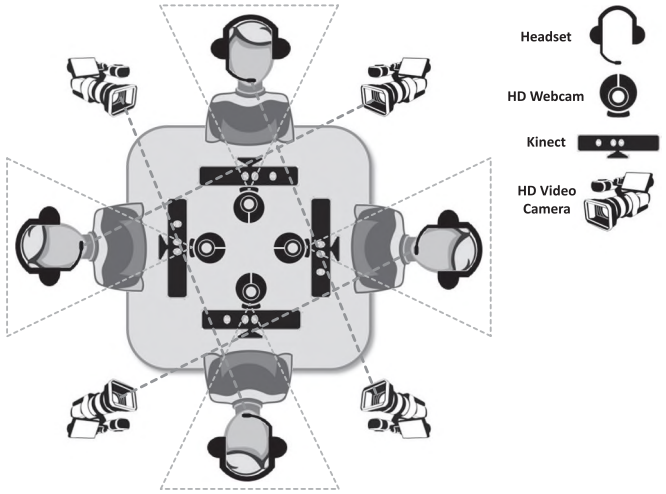


Fig. 12. Round-table collocation of participants during storytelling sessions, including positioning of HD webcams, Microsoft Kinect™ devices, HD video cameras and head mounted microphones.

task is based on the 16 Enjoyable Emotions Induction Task [16]. 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 storytelling sessions the participants were seated in comfortable chairs around a central table, and each participant wore a head-mounted microphone to capture high quality audio recordings. Video signals were recorded using HD webcams. Kinect motion capture technology was used to capture facial features, gaze direction and depth information (see Fig. 12). Synchronisation was achieved using the Social Signal Interpretation framework(SSI) [38]. Participants took turns at recalling a story associated with each enjoyable emotion. The list of enjoyable emotions was randomised for each story telling session, and all of the participants told stories associated with the same emotion in each round of stories. The amount of enjoyment varied depending on which emotion was being recalled and the nature of the story that was being recounted. The story-telling events occasionally evolved into an open discussion, which further facilitated episodes of laughter.

Annotations within the Belfast Storytelling Database [22] segment the audiovisual data on a number of different levels (Fig. 2). 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.

Given the naturalistic, multi-person Belfast Storytelling Database [22] and annotations for smiles, laughs and enjoyment episodes (Section 4), we can carry out a practical comparison study for the classification methodologies and fusion systems that have been discussed so far (Section 3). 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

TABLE 1
Overview of Feature Extraction Methods Applied to Modalities Audio and Video in the Comparison Study

Modality	Channels	short-term feature	long-term feature	total
Audio	Mono Audio, 48 kHz	Pitch, Energy, MFCCs, Spectral, Voice quality	Mean, Median, Maximum, Minimum, Variance, Median, Lower/Upper Quartile, Absolute/Quartile Range	1,451
Video	Action Units, 25 Hz	Upper Lip Raiser, Jaw Lowerer, Lip Stretcher, Brow Lowerer, Lip Corner Depressor, Outer Brow Raiser	Mean, Energy, Standard Deviation, Maximum, Minimum, Range	36

From the mono audio channel and action units captured by Microsoft Kinect™we extract short-term features and compute statistical long-term features from these respectively.

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 percent 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 true real-time system the evaluation has been carried out for the full recordings, i.e., no frames were excluded at any time. Consequently, in case of synchronous fusion a decision had to be forced even for frames where no signal was detected (i.e., no face tracked or silence in the audio channel). We decided to map those frames onto the class with the highest a priori probability (i.e., no-Enjoyment).

5 Comparison Study

Every affect recognition system described in this article uses the same set of features for the audio and video modality, which are described in detail in Table 1: For audio analysis we only 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. As a feature set for characterizing the raw audio streams, we use 1,451 statistical prosodic EmoVoice features described in [37]. Recognizers for video classification are trained with 36 features, gained from statistics over action units provided by the Microsoft Kinect™: i.e., we take the measured activation values of the six action units given by Kinect™tracking for each video frame at 25 fps (upper lip raiser, jaw lowerer, lip stretcher, brow lowerer, lip corner depressor and outer brow raiser) and calculate six statistical measurements (mean, energy, standard deviation, maximum, minimum and range) over a sliding window of 400 milliseconds with 600 milliseconds delta size. This way we obtain 36 features (Table 1) as input for video classification. Input and target features are standardized to zero mean and unit variance on the training set. In the following section, we will discuss the results generated by the various presented approaches to affect recognition (Section 3). 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 the 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. We begin the practical study with the analysis of unimodal classification accuracies which can then be used as a comparison baseline of the multimodal effect of the implemented fusion schemes.

TABLE 2
Unimodal and Synchronous Classification Results

	Synchronous Recognition	
	Audio	Video
Enjoyment	50.16%	67.18%
¬ Enjoyment	60.45%	76.29%
Average	**55.31%**	**71.74%**

The video modality corresponds better to the progression of enjoyment episodes than the audio channel.

The evaluated systems can be applied to real-time applications via the Social Signal Interpretation framework[1] (SSI) [38]. Event based fusion is a native part of the framework, for asynchronous fusion we integrated the CURRENNT[2] library [39], which is originally available as a command line tool for offline training and evaluation. The neural networks can be trained and used within SSI as part of an online recognition system.

5.1 Unimodal Recognition

Table 2 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 [4] for direct framewise classification of enjoyment episodes. Recognition of enjoyment via the audio modality is close to random (55.31 percent). 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 (Fig. 2). With an unweighted 71.74 percent, 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.

The next possible approach-without applying a multimodal fusion strategy-is to consider the temporal flow of

1. http://openssi.net/
2. https://sourceforge.net/projects/currennt/

TABLE 3
Unimodal, Asynchronous Enjoyment Classification Results

	Asynchronous Recognition	
	Audio	Video
Enjoyment	65.41%	76.31%
¬ Enjoyment	75.62%	74.89%
Average	**70.52%**	**75.60%**

The consideration of the temporal flow of observed frames greatly increases classification accuracies.

observed frames. We use BLSTM-NNs (Section 3.2) to model a memory of surrounding frames and therefore incorporate temporal alignments into the classification (Table 8 shows network architectures for audio and video). 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 (Tables 2 and 3). Especially on the audio modality an impressive improvement of 15.21 percent can be observed. The problem of audible cues not fitting well the boundaries of enjoyment episodes is reduced by asynchronous classification.

Table 4 shows the recognition accuracy for laugh and smile events, that can be used to algorithmically derive long-term enjoyment episodes on event level. We use SVM classification for the task of event recognition, as these short term cues should be identifiable within a single frame. 84.05 percent accuracy for audible laughs and 78.98 percent 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 versus 55.31 and 70.52 percent, respectively). Consequently, fusion approaches that are designed to make use of event recognition should be able to utilize audible information to the fullest.

Lastly, we are able to apply event-based recognition approaches to one single modality, by relating recognized events of a single channel back to whole affective episodes via the presented event driven fusion schemes. 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 unimodal scenario (Table 5). Deriving sought affective episodes from audible laughs with the vector fusion approach, we are able to achieve an accuracy of 74.70 percent

TABLE 4
Results for Recognition of Enjoyment Indicating Events Laugh and Smile

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

Short-termed events are easier to classify than the abstract affective target class.

TABLE 5
Event (Vector) Fusion Algorithm Applied to Unimodal Events

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

The indirect approach already results in improved enjoyment classification accuracies, without taking multimodal information into account.

for classification of enjoyment frames. As discussed before, the audio channel can apparently best be employed on the event level. Taking only smile events from the video channel into account results in a recognition rate of 75.76 percent.

5.2 Multimodal Recognition

So far, all 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 are based on the combined insights gained from multiple channels. We will first start with synchronous fusion schemes, the simplest and most common approaches to multimodal fusion. Asynchronous fusion systems apply classification models that are suited to better catch temporal alignments between observed modalities 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.

5.2.1 Synchronous Fusion

Feature-, decision- and model-fusion are obvious approaches to combine multimodal 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 multimodal 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 model level have been tested with very close average recognition rates. Presented results are generated with the product rule (decision level) and stacking (model level)-as described in Section 3.1- and synchronous SVM classification, in order to fully exclude the temporal aspect.

Discrepancies between enjoyment classification on the audio and video modality (Table 2) pass on to these simple synchronous fusion approaches: Feature-, decision-, and model-level fusion perform on an intermediate level between the merged modalities (61.11, 65.86, and 62.17 percent). This is to be expected, as the problematic enjoyment classification models trained on the vocal modality fully contribute to the fusion result.

5.2.2 Direct Asynchronous Fusion

The asynchronous fusion algorithms described in Section 3 roughly use a direct feature fusion approach to affect recognition. But instead of synchronously considering the multiple

TABLE 6
Results for Synchronous Framewise Fusion of Direct Enjoyment Classification Results of the Audio and Video Modality on Several Fusion Levels

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

Poor results of enjoyment classification of the audio channel fully contribute to the fusion result.

channels, the inherent logic of the classification schemes is able to catch asynchronous relationships between modalities. By including multimodal information into the asynchronous classification schemes, the characteristics of an enjoyment episode can be adequately modelled. 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 across the observed modalities and result in enjoyment classification rates of up to 75.76 percent.

We used the same scheme to train deep feedforward neural networks with rectifier neural units and BLSTM-NNs unless described otherwise. All neural networks were trained using the stochastic gradient descent (SGD) algorithm with a momentum value of 0.9. We used grid search to tune the learning rate (1e-1, 1e-2, 1e-3, 1e-4), the number of hidden layers (1, 2, 3), and the number of hidden nodes (16, 32, 64, 128, 256, 512, 1024) (Table 8). Each hidden layer has the same number of hidden nodes. In order to determine the needed parameters, we held out 9 311 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 chose the networks that achieved the best validation error. The validation set was then combined with the training set and we retrained the network on the combined data. The way of combining the validation set was chosen because we found that the networks often benefits from the larger amount of samples. For audio-related features, principle component analysis (PCA) is further used to reduce the dimensionality of the merged feature vector so as to greatly reduce the high computational cost. For PCA, we retain 95 percent of the variance, resulting in an input dimension of 403 for all used networks. 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 Deep-NNs and BLSTM-NNs using the open-source software Theano [2] and CURRENNT [39].

TABLE 7
Asynchronous Fusion Results of the Direct Recognition Approach

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

The memory capabilities of respective algorithms enable the capture of temporal dependencies between observed channels.

TABLE 8
Input Dimension After Principal Component Analysis (PCA) as Well as Optimal Number of Hidden Layers (1, 2, 3), and Corresponding Number of Hidden Nodes (16, 32, 64, 128, 256, 512, 1024) on Training Data After Grid Search

	Applied Neural Networks		
Approach	Input Dim	Hidden Layers	Nodes
Audio			
BLSTM-NN	392	2	16
Video			
BLSTM-NN	36	2	16
Multimodal			
Deep-NN	403	3	1,024
LSTM-NN	403	3	64
BLSTM-NN	403	3	128

5.2.3 Event-Based Fusion

Bringing together the recognition of enjoyment indicating short-term events and the possibility to temporally relate these multimodal 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 percent. 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) are significant. Table 9 also shows a well balanced distribution of accuracies among classes (78.52 percent for Enjoyment and 80.51 percent 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, Fig. 13 plots the confidence values of event recognizers against the correctness of the estimation. Prediction behaviours of modalities resemble each other clearly.

Optimal configuration of parameters have been empirically determined by systematically testing a large number of combinations (Figs. 14 and 15) during the training phase: The most important parameter for vector fusion is the decay speed of registered events. Speed of smile events

TABLE 9
Combination of Event Recognition and the Algorithmic Capture of Temporal Dependencies Between Modalities Leads to Best Enjoyment Recognition Results Measured within the Study at Hand

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

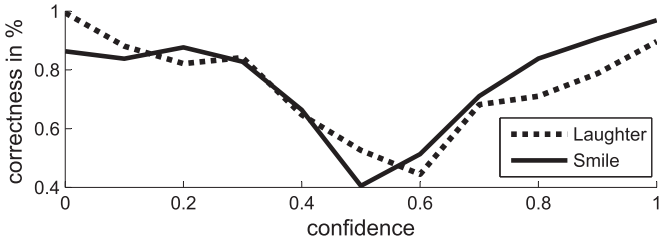


Fig. 13. Frequency of correctly classified frames according to laugh/smile confidence. Similar prediction behaviour allows to directly combine confidence values during the fusion process.

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.

5.3 Summary of Results

In Table 10 we give a summary of the found results in the presented study on naturalistic data. Synchronous unimodal classification results achieved with features from the video modality are used as a baseline-as direct enjoyment classification on the audio channel yields very low accuracy (Table 2). A first improvement (3.86 percent) can be achieved if we switch to asynchronous classification of enjoyment on the video channel (Table 3). The consideration of temporal alignments together with good recognition rates for event recognition (Table 4) leads to an even more accurate classification rate for the event-based approach (Table 5). By combining smile events found in the video modality over time with the vector fusion algorithm, the unimodal result can be raised by 4.02 percent. Note that the term *fusion* does completely apply in this case as we use it mainly for the combination of several modalities.

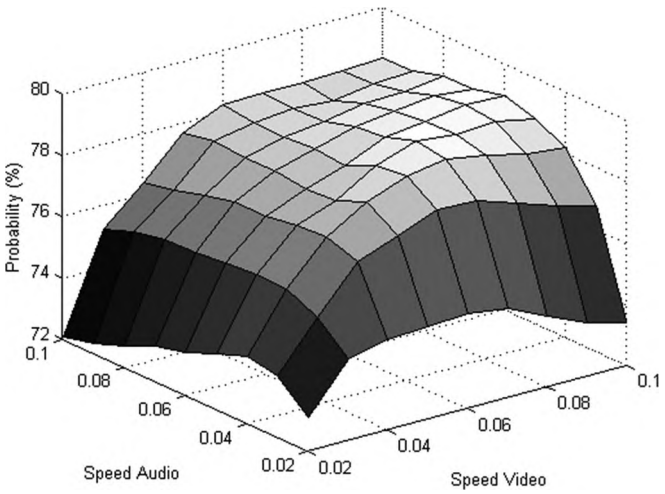


Fig. 14. Influence of audio and video event decay speed on vector fusion performance.

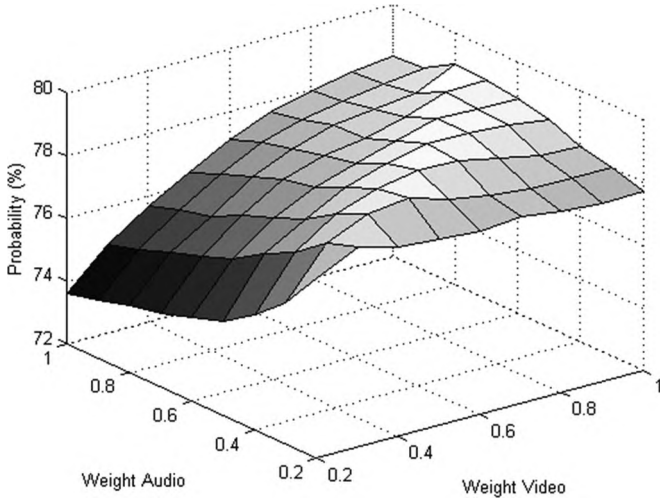


Fig. 15. Influence of audio and video weights on gravity fusion performance. Stable performance is observed if audio and video events are weighted in a ratio of 8 to 10.

First insights gained for unimodal classification on the better suited video channel carry over to multimodal fusion schemes. We first tested common, synchronous feature-, decision-, and model-level fusion strategies to combine information from multiple channels. On average, however, these algorithms lead to a performance which lies between the classification rates of the single channels (Table 6). Because of the low performance of the audio modality, the average accuracy lies 5.88 percent below the unimodal baseline. The neural network based asynchronous fusion systems stick to direct classification, but introduce memory cells to model the asynchronous dependencies between modalities (Table 7). These dedicated approaches are able to better catch the multimodal characteristics of enjoyment episodes and therefore yield recognition enhancements of up to 4.02 percent. Event-based fusion schemes combine an indirect classification approach with algorithms able to leverage the temporal relations between the recognized multimodal events (Table 9). This combination leads to an improvement of recognition accuracy for enjoyment frames of 7.83 percent with the gravity fusion algorithm over unimodal classification.

TABLE 10
Comparison of Tested Approaches to Affect Recognition, Differentiated in Relation to the Classification of Models in Fig. 4

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

Results of respective best performing algorithms are shown with the achieved effect in relation to unimodal, synchronous enjoyment classification on the video channel.

6 Conclusion

Affect recognition systems apply multimodal fusion under the reasonable assumption that combination of information from several modalities improves classification accuracy. However, studies over the last years have shown that the real enhancements of fusion systems compared to unimodal classification are-to say the least-unstable. If we look at the common synchronous fusion approaches (Table 6), 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 multimodal fusion in this case. Fortunately 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 unimodal 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 approaches.

To investigate the above-mentioned options for improvement, we started from the synchronous classification of enjoyment episodes from the audio and video modality. While enjoyment classification on the basis of Kinect facial features showed an acceptable recognition accuracy of 71.74 percent, the audio modality stayed on a close to random level of 55.31 percent. This shortcoming directly influences the performance of the widely used synchronous fusion approaches and results in a mediocre recognition accuracy of 65.86 percent at the decision level. By applying asynchronous classification techniques via neural networks with memory capabilities, these results can be raised to 75.60 percent for the video modality and 75.76 percent for audio-visual fusion systems. Furthermore, the recognition of short-term events as indicators of the target classes turned out to be very reliable with 84.05 percent for voiced laughs and 78.98 percent for visual smiles. Based on this observation, we combined the asynchronous treatment of data with the technique of event classification and event-based fusion. This way, we reached a recognition accuracy of 79.51 percent, which means an improvement of 7.83 percent compared to the unimodal base system and a 13.71 percent higher accuracy than the badly performing synchronous fusion strategies.

Inclusion of temporal observations as well as indirect classification via event recognition have proven to enhance the performance of classification systems and therefore both techniques can be advised to be applied in future affect recognition systems. Event-based fusion strategies applying these techniques also fulfil the requirements demanded by latest considerations about innovative fusion systems [12]: 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.

7 Future Work

Based on the study at hand, various opportunities for future enhancements can be considered or are currently being investigated. The event-based fusion approaches within the study relied on unimodal event recognition with Support Vector Machines. As the ability to consider past frames has proven to give recurrent neural networks a clear edge in recognition accuracy in comparison to conventional classifiers (e.g., Tables 2 and 3), it will be an obvious chance for improvement to use these networks for event recognition. Background knowledge is needed to identify reasonable types of indicator events for a target class and the robust recognition of these is key because errors in event recognition directly map into the event-based fusion result. The combination of picking a good selection of indicator events and the best machine learning techniques to recognize them is needed to further improve the promising results we have seen in this study. Deep learning is another advantage of neural networks, as it skips the problem of feature engineering for a given recognition problem and therefore bears the potential to simplify the addition of modalities and indicator events. The more affective cues event-based fusion algorithms can rely on, the broader the emotional spectrum that can be covered. Using events as meaningful behavioral units allows us to consider more complex and subtle emotional states. For example, a smile is not always a sign of enjoyment, but could also indicate embarrassment-in particular in cases where the gaze is averted from the interlocutor [17]. By considering gaze aversion as an additional event, such distinctions could be captured. In this case, the gaze event should not increase the evidence for enjoyment, but modify the result of the fusion process. To handle such situations, the fusion approach needs to take into account that events are not always used in a redundant manner, but may also complement or even conflict with each other. A promising avenue for future research might be to research to what extent techniques from semantic fusion may be adopted to exploit semantic relationships between events [1].

Finally, we would like to note that the event-based fusion approach is not bound to a particular representation of affective states. However, the results of the experiments demonstrate the potential of an intermediate layer of representation in terms of meaningful events as indicators of affective states. Events can be either mapped onto emotional categories (as in Section 3.3.1) or continuous emotional states in a dimensional space (as in Sections 3.3.2 and 3.3.3). We are currently increasing the repertoire of events in order to cover a broader range of emotion categories and a larger area of the valence-arousal emotion space.

Acknowledgments

The work described in this article received funding from the European Union's Horizon 2020 research and innovation programme (Projects ARIA-VALUSPA, grant agreement no. 645378 and KRISTINA, grant agreement no. 645012).

References

[1] E. André, J.-C. Martin, F. Lingenfelser, and J. Wagner, "Multimodal fusion in human-agent dialogue," in *Proc. Coverbal Synchrony Human-Mach. Interaction*, 2014, pp. 387–410.

[2] J. Bergstra, et al., "Theano: A CPU and GPU math expression compiler," in *Proc. 9th Python Science Conf.*, 2010, pp. 1–7.

[3] R. Brueckner and B. Schuller, "Using deep BLSTM recurrent neural networks," in *Proc. IEEE Int. Acoustics Speech Signal Process.*, 2014, pp. 4856–4860.

[4] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, 2011, pp. 1–27.

[5] S. D'Mello and J. Kory, "Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proc. 14th ACM Int. Conf. Multimodal Interaction*, 2012, pp. 31–38.

[6] R. Duin and D. Tax, "Experiments with classifier combining rules, *Lecture Notes Comput. Sci.*, vol. 1857, pp. 16–29, 2000.

[7] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition, *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[8] P. Ekman, "Basic emotions," *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power eds. Sussex, U.K.: Wiley, 1999.

[9] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE—The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[10] 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, Jun. 2005.

[11] M. Glodek, M. Schels, G. Palm, and F. Schwenker, "Multi-modal fusion based on classifiers using reject options and Markov fusion networks," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 1084–1087.

[12] M. Glodek, et al., "Fusion paradigms in cognitive technical systems for human-computer interaction," *Neurocomputing*, vol. 161, 2015, pp. 17–37.

[13] S. W. Gilroy, et al., "PAD-based multimodal affective fusion," in *Proc. 3rd Int. Conf. Affective Comput. Intell. Interaction Workshops*, 2009, pp. 1–8.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* vol. 9, no. 8, pp. 1735–1780, 1997.

[15] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," *A Field Guide to Dynamical Recurrent Neural Networks*, S. C. Kremer and J. F. Kolen, Eds. Piscataway, NJ, USA: IEEE Press, 2001, pp. 1–15.

[16] J. Hofman, F. Stoffel, A. Weber, and T. Platt, "The 16 enjoyable emotions induction task," Unpublished Research instrument, Department of Psychology, University of Zurich, Switzerland, 2012.

[17] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame," *J. Personality Social Psychology*, vol. 68, no. 3, pp. 441–454, 1995.

[18] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, Feb. 2002.

[19] L. I. Kuncheva, "Switching between selection and fusion in combining classifiers: An experiment," *IEEE Trans. Syst. Man Cybern., Part B: Cybern.* vol. 32, no. 2, pp. 146–156, Apr. 2002.

[20] F. Lingenfelser, J. Wagner, and E. André, "A systematic discussion of fusion techniques for multi-modal affect recognition tasks," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 19–26.

[21] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran, "An event driven fusion approach for enjoyment recognition in real-time," in *Proc. 18th ACM Int. Conf. Multimedia*, 2014, pp. 377–386.

[22] 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 *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 166–172.

[23] 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.

[24] M. Mortillaro, B. Meuleman, and K. Scherer, "Advocating a componential appraisal model to guide emotion recognition," *Int. J. Synthetic Emotions* vol. 3, no. 1, pp. 18–32, 2012.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[26] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. IEEE Int. Acoustics Speech Signal Process.*, 2002, pp. 2013–2016.

[27] M. A. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 3695–3699.

[28] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affective Comput.* vol. 2, no. 2, pp. 92–105, Mar.-Jun. 2011.

[29] R. Niewiadomski, et al., "Laugh-aware virtual agent and its impact on user amusement," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2013, pp. 619–626.

[30] R. Ploikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Magaz.*, vol. 6, no. 3, pp. 21–45, Jul.–Sep. 2006.

[31] F. Rosenblatt, *Principles of Neurodynamics.* New York, NY, USA: Spartan, 1963.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representation by Error Propagation.* Cambridge, MA, USA: MIT Press, 1986.

[33] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[34] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition—A new approach," *in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 1020–1025.

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

[36] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *J. Artif. Intell. Res.*, vol. 10, pp. 115–124, 1999.

[37] T. Vogt, E. André, and N. Bee, "EmoVoice—A framework for online recognition of emotions from voice," in *Proc. 4th IEEE Tutorial Res. Workshop Perception Interactive Technol. Speech-Based Syst.: Perception Multimodal Dialogue Syst.*, 2008, pp. 188–199.

[38] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The Social Signal Interpretation (SSI) framework: Multimodal signal processing in real-time," in *Proc. 18th ACM Int. Conf. Multimedia*, 2013, pp. 831–834.

[39] F. Weninger, J. Bergman, and B. Schuller, "Introducing CURRENNT—the Munich open-source CUDA recurrent neural network toolkit," *J. Mach. Learn. Res.*, vol. 16, pp. 547–551, 2014.

[40] M. Wöllmer, M. Ilmer, 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, pp. 366–380, 2009.

[41] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening," *IEEE J. Select. Topics Signal Process.*, vol. 4, no. 5, pp. 867–881, Oct. 2010.

[42] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audio-visual affective expression recognition through multistream fused HMM," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 570–577, Jun. 2008.

[43] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. 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, Jan. 2009.

Florian Lingenfelser received the MSc degree in informatics and multimedia from the University of Augsburg, Germany, in 2009. He is working toward the PhD degree from University of Augsburg. In 2010 he joined the chair for Human Centered Multimedia of the same University. Among other European projects, he is currently contributing to multimodal data fusion within the Aria Valuspa project and is developing a general framework for the integration of multiple sensors into multimedia applications called Social Signal Interpretation (SSI).

Johannes Wagner received the master of science degree in informatics and multimedia from the University of Augsburg, Germany, in 2007. He is currently employed as a research assistant in the lab of Human Centered Multimedia (HCM) and has been working in several European projects including Humaine, Callas, Ilhaire and CEEDs. His main research focus is the integration of social signal processing (SSP) in real-life applications. He is the founder of the Social Signal Interpretation (SSI) framework, a general framework for the integration of multiple sensors into multimedia applications.

Jun Deng received the bachelor's degree in electronic and information engineering from Harbin Engineering University, in 2009 and the master's degree in information and communication engineering from Harbin Institute of Technology (HIT), Heilongjiang/China, in 2011. He is currently working toward the PhD degree in the MISP Group, TUM in Munich/Germany. His interests include machine learning methods such as transfer learning with an application preference to emotion recognition in speech.

Raymond Brueckner received the BEng and MEng degrees in electronic engineering from the Friedrich-Alexander University Erlangen-Nuernberg, Germany, in 1998. He is currently working toward the PhD degree in the Machine Intelligence and Signal Processing Group, MMK, Technical University Munich, Germany. He is a principal research scientist with Nuance Communications Inc.-Dragon ASR Research Group. From 1998 to 2003 he was part of the speech research group with Ericsson, Germany. From 2003 to 2009 he was a member of the Harman Automotive Systems Speech Group, Germany. In 2009 he joined SVOX AG as a senior research scientist, which got aquired in 2011 by Nuance, for which he has been working since. His current research focuses on machine learning techniques for paralinguistic and emotion recognition.

Björn Schuller received the diploma in 1999, the doctoral degree in automatic speech and emotion recognition, in 2006, and the habilitation and adjunct teaching professorship in the subject area of signal processing and machine intelligence, in 2012, all in electrical engineering and information technology from TUM, Munich, Germany. He is an associate in the Swiss Center for Affective Sciences, University of Geneva, and a senior lecturer of machine learning in the Department of Computing, Imperial College London/UK. He is president of the Association for the Advancement of Affective Computing, elected member of the IEEE Speech and Language Processing Technical Committee, and (co-)authored 5 books and more than 450 publications in peer reviewed books, journals, and conference proceedings leading to more than 8 000 citations (h-index = 44). He is a member of the ACM, the IEEE and the ISCA.

Elisabeth André is a full professor of computer science with Augsburg University and chair of the Laboratory for Human-Centered Multimedia. Prior to that, she worked as a principal researcher with DFKI GmbH where she has been leading various academic and industrial projects in the area of intelligent user interfaces. In summer 2007 Elisabeth André was nominated fellow of the Alcatel-Lucent Foundation for Communications Research. In 2010, she was elected a member of the prestigious German Academy of Sciences Leopoldina and the Academy of Europe. She is also a fellow of the European Coordinating Committee for Artificial Intelligence. Her research interests include affective computing, intelligent multimedia interfaces, and embodied agents.