An Event Driven Fusion Approach for Enjoyment Recognition in Real-time

Florian Lingenfelser
Human Centered Multimedia
University of Augsburg
lingenfelser@hcm-lab.de

Johannes Wagner
Human Centered Multimedia
University of Augsburg
wagner@hcm-lab.de

Elisabeth André
Human Centered Multimedia
University of Augsburg
andre@hcm-lab.de

Gary McKeown
School of Psychology
Queen's University Belfast
g.mckeown@qub.ac.uk

Will Curran
School of Psychology
Queen's University Belfast
w.curran@qub.ac.uk

ABSTRACT

Social signals and interpretation of carried information is of high importance in Human Computer Interaction. Often used for affect recognition, the cues within these signals are displayed in various modalities. Fusion of multimodal signals is a natural and interesting way to improve automatic classification of emotions transported in social signals. Throughout most present studies, uni-modal affect recognition as well as multi-modal fusion, decisions are forced for fixed annotation segments across all modalities. In this paper, we investigate the less prevalent approach of event driven fusion, which indirectly accumulates asynchronous events in all modalities for final predictions. We present a fusion approach, handling short-timed events in a vector space, which is of special interest for real-time applications. We compare results of segmentation based unimodal classification and fusion schemes to the event driven fusion approach. The evaluation is carried out via detection of enjoyment-episodes within the audiovisual Belfast Story-Telling Corpus.

Categories and Subject Descriptors

I.5 [**PATTERN RECOGNITION**]: Applications—*Computer vision; Waveform analysis; Signal processing*

Keywords

affect recognition; social signal processing; multi–modal fusion; event–driven fusion

1. INTRODUCTION

Affective states of human beings refer to the experience of feelings or emotions. These conditions are expressed by the experiencing person through various channels and can be naturally understood by other humans. One of the goals of human computer interaction (HCI) is to automate this perception process and give machines the capability to assess affective states of users [18, 24]. Through the appliance of appropriate sensor technologies and recognition techniques, multi-modal cues that point to certain affective states can be measured and recognized. These sources of evidence lead to the automatic classification of human emotions.

Since emotions are generally observable in multiple channels, the obvious approach is to incorporate as much multimodal information as possible in the classification process [30]. Meaningful features and hints of monitored signals are then to be combined through fusion strategies in order to generate a final prediction. For example, if confronted with a classical audio-visual emotion recognition problem, one would probably first have a look at single spoken sentences, calculate prosodic features from the audio signal and classify the whole sentence with a statistical feature set. In order to enrich descriptive information about the observed signal segment and assumably enhance recognition performance, additional descriptors can be extracted from the video images that are recorded during the spoken sentence in the audio signal. At this point fusion of information from both modalities has to be applied. In the simplest case the two feature sets are merged to be used by a single classifier [26]. More elaborate ways of fusing multiple modalities are in use throughout many affect recognition studies [22, 31, 29, 28, 16, 17]. These fusion strategies and their effects on recognition accuracies will be discussed in the following sections. Some facts can however be anticipated here: Present studies have shown varying degrees of success or even failure of classical fusion approaches [2]. Assuming decent data content in all considered modalities, one would generally expect a steady classification gain from adding additional information.

A possible reason for the unsteady performance of presented fusion schemes could lie in the initiation of consideration of multiple signals and appliance of the above mentioned fusion algorithms. In off-line studies, the triggering of fusion processes is simply given by the annotation boundaries; in a real-time scenario this is typically done by detecting the on and offset of a relevant time-interval in one modality. Afterwards fusion techniques are called for classification throughout all available modalities. Consequently, the seg-

MM'14, November 3–7, 2014, Orlando, FL, USA
Copyright 2014 ACM 978-1-4503-3063-3/14/11$15.00.
http://dx.doi.org/10.1145/2647868.2654924.

mentation of a cue in one modality is forced upon other available channels. What if nothing is happening in the face 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. Cutting fixed segments through multi-layered signals does seem to be undesirable. One could think of adding deltas to additionally concerned modalities, but this approach would most likely lead to a hard-wired construct, that is hard to define and not generalizable. So how do we solve the problem of non-aligned cues in multiple signals?

A first step is to reject the assumption that all relevant cues happen at the same time in all modalities. Practical observations demonstrate the need to detect events for each modality separately. But if all signal–events are treated individually, we have to find ways to relate them for proper fusion of identified information. Therefore, we introduce an event driven fusion model that incorporates the temporal relation between the events. It is specially suited for real-time applications. An early prototype has been already successfully applied at the eNTERFACE 13 workshop in Lisbon[13]. For evaluation we will analyse a suitable corpus for recognition of enjoyable emotions.We compare single modality accuracy to segmentation based fusion approaches and event driven fusion - concerning framewise classification of enjoyment. We define enjoyment as an episode of positive emotion, indicated by visual and auditory cues of enjoyment, such as smiles and voiced laughters. This enables us to compare overarching enjoyment–annotations to accumulated indication-events. Based on this evaluation we can finally discuss advantages and disadvantages of event driven fusion for affect recognition.

2. MULTI-MODAL FUSION

In a standard classification task, recognizers are trained with samples of pre-segmented data. This segmentation is achieved by annotation of the recorded data. Experts review the data, marking time-segments of interest and providing them with a pre-defined label that describes the nature of the respective time period. Resulting data samples are then subject to feature extraction techniques. When dealing with one modality, this procedure is carried out on one specific kind of signal and one classification model can be trained with the resulting samples. In multi-modal classification, every single signal needs an adapted feature extraction step, resulting in feature sets for every observed signal. Reasonable combination of available information is the challenge of multi-modal fusion approaches.

2.1 Segmentation–Based Fusion Approach

When confronted with fusion of multiple signals, a vast amount of eligible strategies come into consideration [20]. Possible methods can be differentiated by the levels at which they are executed. Authors in [30] cite 18 studies dealing with audio-visual fusion. Here, they distinguish between feature-, decision- and model-level fusion.

A very straightforward way to fuse all observed modalities is to merge all calculated features into a single and high dimensional feature set for one single classification model (feature level fusion). The accumulated features contain a greater amount of information than a single modality. Prediction based fusion, as proposed in [19], tries to discriminate classes by modelling spatial and temporal relationships between multi-modal features. Decision level fusion sums up combination 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, look-up tables and algebraic combination rules such as sum rule or product rule). Feature and decision level fusion include the most standard approaches used in most studies concerning multi-modal fusion experiments. In model level fusion (e. g. stacked generalisation [23] 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.

A fair amount of studies incorporate these segmentation based fusion techniques for combination of observed signals and final classification. Meta studies like [2] compare gathered results and give an overview: Some report remarkable accuracy gains over uni-modal classification, others do not notice statistically relevant benefits. Even substantial drops in classification quality are sometimes registered. In numbers, the effect of multi-modal fusion in comparison to uni-modal classification range from a +27.4% gain in recognition performance to a -9.0% drop in overall accuracy for a total of 30 compared studies. The study also points out, that classification improvements are far more likely to be achieved on acted data than on natural or semi-natural recordings. Such meta comparisons do not go into detail about the applied fusion schemes. Studies like [3, 11, 10, 6] examine rather basic fusion strategies and sometimes advise on which scheme dominates others. Results are not consistent throughout mentioned experiments. Furthermore, 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).

2.2 Asynchronous Fusion Approach

In segmentation based fusion approaches, analysis of all modalities is initiated and margined by a comprehensive annotation for the given classification problem [12]. For example, when doing audiovisual emotion recognition it is a common strategy to trigger analysis of further modalities by voice activity detection in the vocal modality. Whenever there is activity in the voice, classification of facial expressions is done during this time segment. Fusion algorithms are then applied to the cues of the extracted time-slice. This approach of triggering multi-modal fusion from a single annotation or modality has at least one severe drawback: Additional cues in further modalities can be expected but are not guaranteed to coexist at the time frames or in the worst case, are not present at all. Imagine the audio-visual affect recognition scenario: The vocal component of an emotional expression may be signalled before the facial component. This way, the observed segment of modalities does indeed fit to relevant data in the audio signal, but boundaries of facial activity are shifted. Such component asynchronicity fully contributes to the multi-modal fusion and negatively influences final classification.

An elegant way of fusing modalities without forcing decisions from all channels in every time slot is offered by dynamic classification. Since dynamic classifiers work on continuous streams of short-term features, it is not necessary to force a fusion decision ”from above”. Instead, they (principally) have the ability to model temporal relations between the streams and learn when and how multimodal information should be combined. Dupont *et. al* [4] 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. Song *et. al* [22] proposed a tripled Hidden Markov Model (THMM), which is able to integrate three or more streams of data and allows the state asynchrony of the sequences while preserving their natural correlation over time. Zeng *et. al* [31] 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), where 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, have also been proposed [15].

To overcome the computational complexity of asynchronous Hidden Markov model (AHMM), Wöllmer *et. al* [29] 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, Artificial Neural Networks (ANN) 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. In fact, LSTM-NNs have been successfully applied 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 [28]. Likewise, in a similar emotion recognition task, this approach successfully fuses facial expressions, shoulder gestures and audio cues [17].

2.3 Event Driven Fusion Approach

While the aforementioned asynchronous fusion approaches theoretically outperform segmentation based schemes, they also have some drawbacks. One severe disadvantage comes from their complexity in terms of training and decision taking. Since it is difficult to understand how the network reaches a decision, applying it in a real-time system bears the risk that the learned model parameters may poorly translate if applied in a possibly less controllable environment. Furthermore, once trained, they function as a black box whose hard-wired parameters leave little opportunity for adjustments to the new conditions. Another issue, which is gladly overlooked in pure offline studies, is the problem of *missing data*. Missing data can occur when either no useful information *can* be detected (e.g. the user is not looking into the camera), or because there *is* nothing useful to detect (e.g. the user is not talking), or last but not least, due to a failure of one of the sensors.

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. Introducing events as an abstract intermediate layer effectively decouples uni-modal processing from the final decision making. 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 of the modalities does not cause the collapse of the whole fusion process. In some sense this kind of event-driven fusion is similar to semantic fusion used to analyse the semantics of multi-modal commands, and typically investigates the combination of gestures and speech in new-generation multi–modal user interfaces [14]. However, only few attempts have been made to apply event-driven concepts for automated emotion detection.

In an artistic Augmented Reality installation, the Callas Emotional Tree [7], Gilroy *et. al* uses event-based fusion to derive the affective state of a user in real-time. The basic idea of their approach was to derive emotional information from different modality-specific sensors and map it onto a continuous affective space spanned by the three dimensions Pleasure, Arousal and Dominance (PAD model). Since the application depended on a continuous assessment of the affective user state, the current state of the fusion system was constantly represented by a vector in the PAD space. And the direction into which the vector would move was set by a bunch of vectors representing the single modality–specific contributions. The values of those guiding vectors was updated whenever a new affective cue was detected or otherwise decayed over time. A different approach for predicting user affect in a continuous dimensional space based on verbal and non-verbal behavioural events (e.g. smiles, head shakes, or laughter), has been published by Eyben *et. al* [5]. In their system events are seen as ”words”, which are joined for each time segment and converted to a feature vector representation through a binary bag–of–words (BOW) approach. Tests on an audiovisual database proved the proposed string-based fusion to be superior over conventional feature-level modelling.

3. EVENT BASED VECTOR FUSION

As implied by first attempts for event driven fusion, a possible way to avoid the problem of restricting segmentations is to have customized annotations for every single modality, from which the recognition of single events – that indicate the sought classification – can be learned. The task of the event driven fusion algorithm then has to be to accumulate these indicating events, take their temporal flow into account and, finally, to classify each single time frame and give boundaries for the recognition on the timescale.

3.1 General Requirements

A real-time event driven fusion scheme, meant to reduce negative effects of the segmentation problem, must meet certain requirements. It should be based on separated event detection in observed signals and its inherent fusion rules must consider the temporal flow of all detected events.

Temporal Component

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 decrease over time - as the moment of occurrence shifts further back in time – until the influential potency reaches a value of zero and 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.

Combining Modality Based Events
The additional effort of 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 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
The result of the fusion scheme is calculated by temporal influences (expressed through momentary weights) of registered events. This result will consist of n continuous confidence values (typically valued and normalized between zero and one) for an n-dimensional classification problem. The continuous fusion result is accessible at any point in time. This circumstance is especially valuable in real–time scenarios, where reactions to changing conditions have to be carried out as fast as possible.

Handling Missing Data
The last demand on the real–time system, is an implementation that resists the temporal absence of cues from one or more modalities. This can result for example from missing activity in a modality, tracking problems, or even the breakdown of attached devices. If recognizers involved in decision making each represent the observations of an associated modality, 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 malfunction is recognized and the corresponding classifier's (most likely counter–productive) contribution is accordingly rated or completely left out of the fusion process.

3.2 Algorithm

The proposed fusion algorithm is based on preceding work done by [8] (section 2.3). We generalize this approach by designing a fusion scheme that operates in a user-defined vector space.

3.2.1 Vector Space and Event Vectors

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 the following parameters:

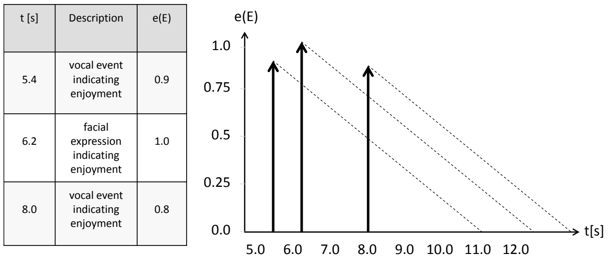
- Confidence Value
  One for each defined axis in the event space. This defines the position of the vector within the dimensional model. For instance, it can be dynamically calculated from the probabilities of a detected cue.

- Vector Weight
  The vector weight is a quantifier for the initial weighting the event has in the calculation of the fusion result. It is defined by the modality the event is detected in and serves as a regulation instrument for emphasizing more reliable information sources. If, for example, one modality is generally better suited for the given classification problem, it can be assigned a higher overall weight. The weight can also be defined by the context. For example, in case of a high noise level, audio might be given less weight.

- Decay Speed
  This is also defined for each modality and describes the average lifespan of cues extracted from the respective signal. If determines the time it takes for the event's influence to decrease to zero and get discarded. Events that strongly indicate the fusion's target class can be given longer decay times, in order to prolong their influence on the result.

In our case, these parameters are empirically determined by systematically testing a large number of combinations for the enjoyment recognition task (see 5.4 for detailed analysis). Figure 1 shows a series of events and their confidence values in the event space. The weight (and therefore influence on the fusion result) of an event vector decrease over time (see dotted line) until the vector is completely removed.

For real-time enjoyment recognition, each frame of audiovisual data undergoes checking for correct face tracking and voice activity. If consequently possible, a SVM classifier for smile recognition (trained with 36 statistical features over action units and smile annotation) and a SVM classifier for laughter recognition (trained with 1451 statistical prosodic features and laughter annotation) each give a normalized probability for the respective event. These confidence values directly map the probabilities given by the SVM models and are used to create uni-modal event-vectors in the multimodal vector space (with resulting event-values in the range of zero to one). Influences (weights) of these events hyperbolically decrease over time, precisely calculated by the initial weight-parameter and speed-parameter of the corresponding modality.

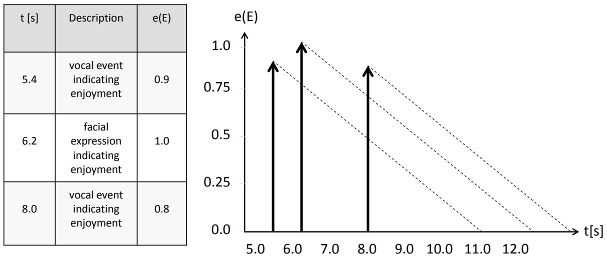


t [s]	Description	e(E)
5.4	vocal event indicating enjoyment	0.9
6.2	facial expression indicating enjoyment	1.0
8.0	vocal event indicating enjoyment	0.8

Figure 1: Multi-modal events mapped into the event space.

3.2.2 Fusion Vector

A mass centre is calculated at each frame for all active (weight greater zero) events, by summing up all event-values modified by their current influence (decreased weight) and averaging over the number of active events. The fusion result itself is a vector, which approaches the calculated mass centre with a predefined speed-parameter (figure 2). If this vector rises above a specified classification threshold, we classify the frame to contain enjoyment. This way, we logically fuse smile and laughter events for enjoyment recognition and can evaluate the fusion result against enjoyment annotations frame by frame.



t [s]	Description	e(E)
5.4	vocal event indicating enjoyment	0.9
6.2	facial expression indicating enjoyment	1.0
8.0	vocal event indicating enjoyment	0.8

Figure 2: The fusion vector (solid line) approaching the temporary mass centre (bars) and then decreasing to a neutral state.

3.2.3 Characteristics

The fact that the fusion vector does not instantly assume the value of the mass centre, 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 vectors have to be adapted to the observed classification problem.

Realizing the premise of separated events turns out to be a labour–intensive task. In practice, it takes considerably more effort to independently identify events in different signals than triggering interpretation in all modalities by a single signal–event. A deeper understanding of every single modality is needed and the signal processing tasks rise proportionally, as meaningful segments now have to be found in each modality. These unrelated segments need to be interpreted and forwarded to the fusion algorithm as events. A strong technical framework for multi–modal event detection in real–time is needed as a foundation for the realisation of an event driven fusion scheme.

4. A FRAMEWORK FOR EVENT DRIVEN FUSION

The Social Signal Processing framework (SSI)[1] [27] offers special support for the development of online recognition systems from multiple sensors. A list of special traits make the SSI framework a good choice for implementing the event driven fusion approach: An architecture is established to handle diverse signals in a coherent way, no matter if it is a waveform, a heart beat signal, or a video image. Live sensor input is available for a long list of hardware devices and new ones can be implemented via offered interfaces. Thereby, implementation details related to real–time processing such as buffering, synchronization, and threading are hidden from the developer of additional content. Components to process captured signals and assemble machine learning pipelines are included in the framework. Possibilities range from real–time signal processing to high-level feature extraction and online classification.

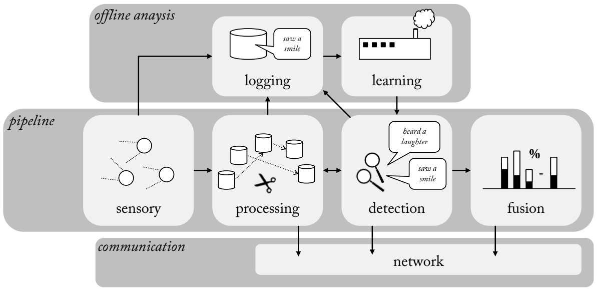


Figure 3: SSI framework: Offline trained recognizers applied in a pipeline for event fusion in real–time.

SSI breaks data handling down to two basic data structures, that perfectly fit the idea of event driven fusion: Streams and events. Data read from a sensor is transformed into a stream, i.e. a continuous flow of samples with fixed sample rate and size. Various transformation algorithms (e. g. filtering, feature calculation, etc.) can be applied to manipulate and further process a raw data stream.

In addition to streams, SSI features the concept of events. Events are meant to describe relevant parts of streams. Single events are usually generated from continuous streams by applying some kind of activity detection. Their length is variable and they may contain additional data, such as a feature set or textual descriptions. Whenever on and offsets are detected, events are sent to the event board. Recognition components that have subscribed to the event are now informed. According to the segment they can request stream chunks or corresponding feature sets and feed them, for example to a single classifier. Classification probabilities are again published as events and can be further processed by event driven fusion schemes.

5. ENJOYMENT RECOGNITION

For evaluation of the event driven fusion approach, we picked the task of enjoyment recognition. We define enjoyment as an episode of enjoyable emotion. These episodes are typically accompanied by visual and auditory cues: We summarize voiced laughters and unvoiced laughters in the audio modality as laughters. The visual component of a laugh as well as visual smiles are in the following denoted as smiles. An annotation of an enjoyment segment will most likely contain one or more of these indicators of enjoyment (figure 4).

5.1 Belfast Story–Telling Corpus

The training corpus for evaluation was taken from the first session of the Belfast Storytelling corpus. The corpus was comprised of six sessions of groups of three or four people telling stories to one another in either English or Spanish. The storytelling task was based on the 16 Enjoyable Emotions Induction Task [9]. 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 session the participants were seated in com-

[1] http://www.openssi.net/

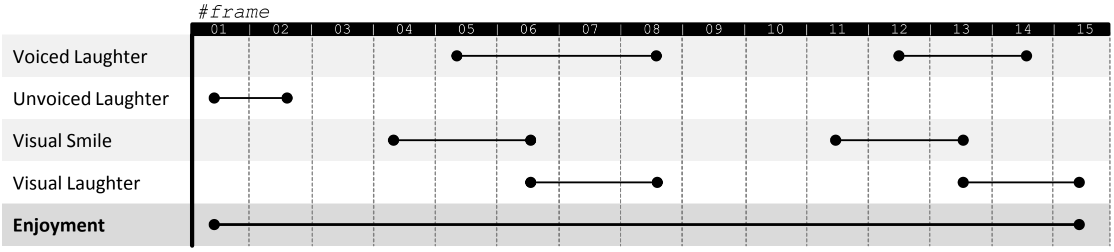


Figure 4: Exemplary annotation of a full enjoyment episode aligned with various voiced and visual cues emitted by the user. For each frame (bordered by dotted lines) a decision has to be made by the fusion system. In a conventional segmentation–based approach each frame is seen in isolation, i.e. a decision 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 2 and 4), some of the frames, which are spanned by the enjoyment episode do actually not overlap with any observable cues (see e.g. frame 9 and 10). Those frames are likely to be misclassified by a segmentation–based approach. The event–driven fusion approach proposed in this paper, which takes in account the temporal asynchronicity of the events, is able to overcome frames with sparse cues of enjoyment based on information of preceding frames.

fortable chairs around a central table, and each participant wore a head-mounted microphone to capture high quality audio recordings. Video signals were recorded using Logitech Pro HD webcams. Kinect motion capture technology was used to capture facial features, gaze direction and depth information. 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 laughter 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. First session involved three male native English speakers recounting stories to one another. Manual annotations of enjoyment episodes and laughter / smile events have been created for this session with the ELAN annotation tool [21] (see figure 4 for an exemplary excerpt).

To capture synchronised data we required the use of 9 computers and a Network Attached Storage (NAS) system. Streaming the visual data from a single participant required a dedicated computer for each HD webcam and Kinect, making a total of 8 computers to capture the data. The audio signals were captured using a ninth computer. The HD Webcams streamed video data to the computers at 25fps, with a resolution of 1024x576 for three of the cameras and 960x720 for the fourth camera. We also used standard video recording equipment as a backup recording system. Webcam streams were compressed with the Huffyuv lossless codec and later compressed using the lossy H264 to make more usable file sizes. The audio from each head mounted microphone was fed into a MOTU 8pre FireWire audio interface preamp, and from there into another computer with Firewire 800 recording hard drives. Audio was recorded using wav format files (mono, 48000Hz, 24-bit PCM). Each session lasted about 120 minutes, resulting in approximately 75 minutes recording time. Synchronized recording of data streams was achieved using the SSI software (section 4).

5.2 Enjoyment Recognition Systems

Figure 4 depicts multiple annotation tracks for the Belfast Story–Telling Corpus. The overarching enjoyment episode includes several segments of smiles and laughters. Training of the following recognition systems is based on these annotations. All classification systems perform recognition on a framewise basis: A decision, if enjoyment is present within the evaluated person or not, is made every 400 milliseconds within a window of one second. Each recognition system is subject independent – two of the persons of session one (section 5.1) are used for training the recognizers needed for compared approaches, one person is used for testing. Given a recording length of approximately one hour per person, this leads to a rough total of 18.000 samples for training and 9.000 samples for testing. Considering class imbalances within testing samples, we use the unweighted average as evaluation criterion.

Uni–Modal Classification
Based on the segmentation given by the annotations, we recognize enjoyment directly from the single modalities audio and video. As feature–sets for characterizing the raw audio streams, we use 1451 statistical prosodic EmoVoice features [25]. Recognizers for video classification are trained with 36 features, gained from statistics over action units provided by the Microsoft Kinect^{TM}. Both feature extraction steps are calculated within the SSI framework (section 4). As computational model for classification, we use LibSVM's support vector machines [1] with a linear kernel.

Segmentation Based Fusion
Segmentation based fusion approaches on the feature, decision and model level (section 2.1) are applied to combine both modalities for direct enjoyment classification (using the same enjoyment annotation track as the uni–modal classification systems). From these experiments we can draw first conclusions if the multi–modal information can deliver classification improvements, if the same annotated time segments are sliced through modalities.

	Uni–Modal Classification		**Segmentation–Based Fusion**		
	Audio	**Video**	**Feature**	**Decision**	**Model**
Enjoyment	50.26%	66.85%	73.75%	76.41%	76.08%
¬ Enjoyment	60.47%	76.90%	66.17%	61.53%	56.04%
UA	**57.14%**	**73.62%**	**68.64%**	**66.39%**	**62.58%**
WA	**55.37%**	**71.88%**	**69.96%**	**68.97%**	**66.06%**

Table 1: Results for uni–modal classification and segmentation based fusion on feature, decision and model level. Trained on overall enjoyment annotations.

Modality-Tailored Fusion
Afterwards we try to recognize the annotated enjoyment segments indirectly from tailored annotations: Instead of using the annotations for whole enjoyment episodes for both modalities, we annotate audible occurrences of laughters within the audio channel and visible laughters and smiles in the video separately. These tailored annotations are then used to train classification models for detecting these enjoyment indicating cues, rather than recognizing enjoyment directly (same feature sets as for enjoyment classification are applied).

Modality-tailored fusion is meant as an intermediate and experimental step, in which these modality tailored cue-recognizers are used directly in decision and model level fusion schemes. The models trained on enjoyment segmentations are therefore replaced, probabilities given to each frame by classifiers meant for detecting audible and visual laughters are mapped to the corresponding enjoyment classes.

Event Driven Vector Fusion
Finally we apply event–based vector fusion to compare this indirect, event–based way of fusing multi–modal information to single channel classification and segmentation based fusion performance. The events have to be detected and generated by the framework. Before classification, activity recognition is performed for each modality, for example testing if there is more than noise in the audio channel. Such pre–processing can introduce additional prediction errors (as the activity recognition is also not always correct), but a very crucial point in robust real–time systems. For this reason we simulate the process also for evaluation. Recognizers consider every frame that pass activity recognition and generate events for smiles and laughters respectively. Confidence values of these events correspond to the recognition probabilities of smile and laughter detectors. All currently active events are finally considered for mass–centre calculation and therefore influence the course of the fusion vector (section 3). For every frame, we calculate the current position of the fusion vector and based on its current position we decide if enjoyment is present or not within the observed time frame.

In order to simulate a true real–time system it should be noted that evaluation has been carried out for the full recordings, i.e. no frames were excluded at any time. Consequently, in case of segmentation-based fusion a decision had to be forced even for frames where no signal was detected (i.e. no face tracked and silence detected in the audio channel). We decided to map those frames onto the class with the highest a priori probability (i.e. no enjoyment).

5.3 Results

Result tables report unweighted (average accuracy across frames) and weighted (average accuracy across classes) recognition results (UA / WA). During the discussion, we will focus on the weighted recognition performance, as classified frames contain less samples of occurring enjoyment as well as audible and visible laughter. Table 1 shows recognition results for single channel classification and segmentation based fusion algorithms that use classification models trained directly on enjoyment annotations. Recognition of enjoyment via the audio modality is close to random (55.37%). Expressive 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 4). With a weighted 71.88%, 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. These discrepancies pass on to segmentation based fusion approaches: Feature, decision and model level fusion perform on an intermediate level between the merged modalities (69.96%, 68.97% and 66.06%)[2]. This is to be expected, as the problematic classification models trained on the vocal modality fully contribute to the fusion result.

	Event Detection		
	Audio	**Video**	
Laughter	76.51%	78.15%	**Smile**
¬ Laughter	91.66%	79.61%	**¬ Smile**
UA	**90.99**	**79.31**	**UA**
WA	**84.09**	**78.88**	**WA**

Table 2: Result for uni–modal event recognizers for laughters and smiles. Trained on modality tailored annotations.

Table 2 gives insight into the capability of event detection recognizers for the audio and video modality. These are not trained with the bi–modal annotations of enjoyment, but with tailored, more narrow uni–modal annotations for actual laughter occurrences and smiles respectively. When looking at laughter classification within the audio channel, it becomes clear that detection of these short indication–

[2] 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 2.1.

	Modality–Tailored Fusion		Event–Driven Fusion
	Decision	**Model**	**Vector Based**
Enjoyment	55.16%	66.75%	76.18%
¬ Enjoyment	90.32%	80.54%	81.37%
UA	**78.84%**	**76.04%**	**79.68%**
WA	**72.74%**	**73.65%**	**78.78%**

Table 3: Results for modality tailored fusion on decision and model level and event driven vector fusion.

events is by far more reliable than recognition of whole enjoyment episodes (84.09% to 55.37%). Differences between smile– and enjoyment recognition on basis of the video channel is not as massive: The low recognition difference of 7.00% (compared to auditory enjoyment and laughters) depicts the high correlation of experienced enjoyment and positive facial expressions.

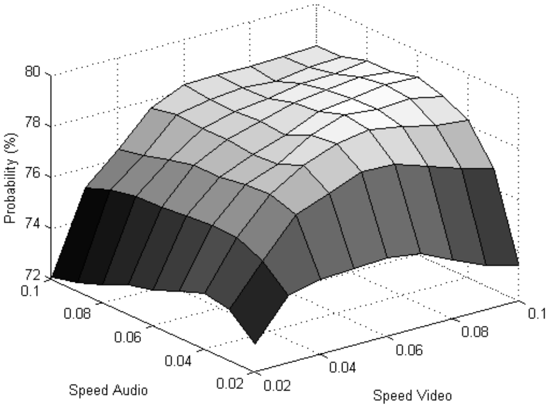


Figure 5: Influence of audio and video decay speed on vector fusion performance.

However, we can not directly use these event recognition models for enjoyment detection. They are meant to be further processed by an event–driven fusion algorithm, such as vector fusion. But first we experiment with an intermediate step: Event recognition models are used in segmentation based fusion algorithms on decision and model level. Laughter and smile detections are simply mapped to the enjoyment class and fed into the fusion process. Results of this procedure are described in the first entries of Table 3 (modality–tailored fusion). Both fusion approaches deliver good results. With a weighted average of 72.74% on decision– and 73.65% on model–level they exceed uni–modal classification of enjoyment on the more qualified video channel (71.88%). By combining 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. Table 3 shows clearly, 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. At this point, event–driven fusion schemes can gain further improvements over previous approaches.

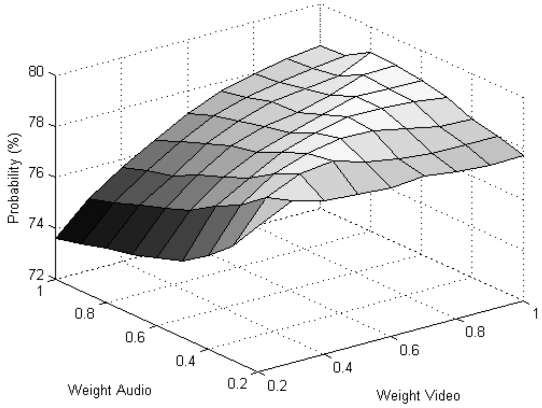


Figure 6: Influence of audio and video weights on vector fusion performance. Stable performance is observed if audio and video events are weighted in a ratio of 8 to 10.

5.4 Parameter Analysis

As described earlier (section 3), the performance of event driven fusion depends on the three parameters confidence, weight and speed of the events. To achieve optimal results, a reasonable configuration of the three parameters has to be found. Confidence is directly derived from probabilities given by the event detectors. This derivation only makes sense if confidence values of given classifiers are comparable. To prove this assumption, Figure 7 plots the confidence values of event detectors against the correctness of the estimation. Prediction behaviours of modalities resemble each other clearly.

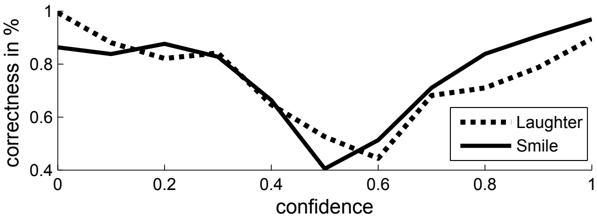


Figure 7: Frequency of correctly classified frames according to laughter / smile confidence. Similar prediction behaviour allows to directly combine confidence values during the fusion process.

Optimal configuration of weight and speed parameters have been empirically determined by systematically testing a large number of combinations (figures 5 and 6): Smile events are weighted with full influence, as the occurrence of smiles correlate well with the boundaries of enjoyment episodes. Their decay speed is regulated high, as the beginning and ending of enjoyment are often similar to presence and absence of smiles in the face and the fusion vector should rise and fall fast whenever smiles are detected or not. The decay speed of laughter events is regulated low. Laughters are considered a strong indicator of enjoyment and whenever they occur we expect the enjoyment episode to last for several frames afterwards. Best performance is achieved if laughters are weighted less than smile events – again due to the fact that smiles better describe the limits of enjoyment segments. The optimal ratio lies around 8 to 10 (figure 6).

Taking these findings and the possibility to temporarily relate the detected events into account, event–driven vector fusion achieves an average recognition rate of 78.78%. This is the best result we were able to achieve for enjoyment recognition during our experiments with examined approaches. According to McNemar's Chi-Squared Test ($p < 0.05$), improvements in comparison to the second best approach (modality tailored fusion on model and decision level) are significant. Table 3 also shows a well balanced distribution of accuracies among classes (76.18% for Enjoyment and 81.37% for ¬Enjoyment), which shows that event driven fusion is accurate in detecting whole episodes of enjoyment and models their boundaries well.

6. CONCLUSION

Affect recognition systems apply multi–modal fusion under the reasonable assumption that combination of information from several modalities does improve classification accuracy. However, studies over the last years have shown that the concrete enhancements of fusion systems compared to uni–modal classification are – to say the least – unstable. A possible problem causing this varying performance is that overarching segmentations for given classification problems are used throughout observed modalities, resulting in segmentation based fusion approaches. In this study we have specified and implemented an event driven real–time fusion system for affect recognition. This kind of approach does not directly fuse identical timeframes throughout modalities, but calculates probabilities indirectly by accumulating shorter, detection–indicating and possibly time–shifted events. This approach demands additional annotation, segmentation and training steps, but our evaluation shows a promising potential of the event driven approach:

Given the affect recognition task of enjoyment classification on the Belfast Story–Telling Corpus, we exemplary compare uni–modal and segmentation based, multi–modal fusion systems to event driven vector vector fusion. While the segmentation for enjoyment episodes indeed fits well for classification via the video modality, it is not very suitable for the audio channel. This fact results in acceptable recognition accuracy when using only the video modality and very bad results for audio classification. Segmentation based fusion aligns between these accuracies and performs on an intermediate level. It then, recognizes the enjoyment–indicating events of laughters and smiles and processes them further with event driven vector fusion. We empirically determined parameters for speed and weight distributions among modalities. Best performance was yielded when smile events were given a higher weight and speed than laughters. Laughters are a strong indicator of enjoyment and should therefore have a long–term influence on fusion; smiles on the other hand need quick reaction time as they describe well the margins of enjoyment. They also profit from higher weightings as this allow smiles to hold steady during longer enjoyment–periods with no vocal activity. Based on this configuration (78.78%), we were able to enhance enjoyment recognition accuracy by 6.09% compared to uni–modal classification (video channel), 8.82% compared to segmentation based fusion (feature level) and 5.13% compared to modality-tailored fusion (model level).

7. FUTURE WORK

Having identified potential problems of segmentation based fusion and the capabilities of an event driven approach, many interesting investigations open up: By now, we have trained few event recognizers for a valence related affect classification problem. The amount of event detectors can be raised by a great amount and expanded to further modalities. Arousal related classification, e.g. on the basis of physiological signals and events, potentially enables recognition coverage of the 2–dimensional valence–arousal emotion space. The presented vector fusion implementation is just one way of relating event detections for multi–modal fusion. Because of its accessible and understandable logic and structure, vector fusion is a very good starting point for analysis of the event driven fusion approach. However, more complicated network structures seem very promising and are to be examined and compared in future studies.

8. ACKNOWLEDGMENTS

The work described in this paper is funded by the European Union under research grant ILHAIRE, a Seventh Framework Programme (FP7/2007-2013) under grant agreement $n°$270780 and CEEDs (FP7-ICT-2009-5).

9. REFERENCES

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[2] S. K. D'Mello and J. Kory. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In L.-P. Morency, D. Bohus, H. K. Aghajan, J. Cassell, A. Nijholt, and J. Epps, editors, *ICMI*, pages 31–38. ACM, 2012.

[3] R. Duin and D. Tax. Experiments with classifier combining rules. In *Lecture Notes in Computer Science*, volume 1857, pages 16–29. Springer, 2000.

[4] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141 – 151, 2000.

[5] F. Eyben, M. Wöllmer, M. F. Valstar, H. Gunes, B. Schuller, and M. Pantic. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, FG 2011*, pages 322–329, USA, March 2011. IEEE Computer Society.

[6] G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):942–956, 2005.

[7] S. W. 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 *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 945–948, New York, NY, USA, 2008. ACM.

[8] S. W. Gilroy, M. Cavazza, M. Niranen, E. Andre, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, and M. Billinghurst. Pad-based multimodal affective fusion. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009.

[9] J. Hofmann, F. Stoffel, A. Weber, and T. Platt. The 16 enjoyable emotions induction task (16-eeit). 2012.

[10] L. I. Kuncheva. Switching between selection and fusion in combining classifiers: An experiment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 32(2):146–156, 2002.

[11] L. I. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286, 2002.

[12] 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 '11, pages 19–26, New York, NY, USA, 2011. ACM.

[13] M. Mancini, L. Ach, E. Bantegnie, T. Baur, N. Berthouze, D. Datta, Y. Ding, S. Dupont, H. Griffin, F. Lingenfelser, R. Niewiadomski, C. Pelachaud, O. Pietquin, B. Piot, J. Urbain, G. Volpe, and J. Wagner. Laugh when you're winning. In C. T. R. J. C.-M. Rybarczyk, Y., editor, *Innovative and Creative Developments in Multimodal Interaction Systems, Proceedings of 9th IFIP WG 5.5 International Summer Workshop on Multimodal Interfaces, eNTERFACE 2013*, volume 425, Lisbon, Portugal, 2014. Springer.

[14] G. U. Mehlmann and E. André. Modeling multimodal integration with event logic charts. In L.-P. Morency, D. Bohus, H. K. Aghajan, J. Cassell, A. Nijholt, and J. Epps, editors, *ICMI*, pages 125–132. ACM, 2012.

[15] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled hmm for audio-visual speech recognition. In *in International Conference on Acoustics, Speech and Signal Processing (CASSP'02)*, pages 2013–2016, 2002.

[16] M. A. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *ICPR*, pages 3695–3699. IEEE, 2010.

[17] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *T. Affective Computing*, 2(2):92–105, 2011.

[18] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. *Human Computing and Machine Understanding of Human Behavior: A Survey*, volume 4451, pages 47–71. 2007.

[19] S. Petridis, M. Pantic, and J. F. Cohn. Prediction-based classification for audiovisual discrimination between laughter and speech. In *Automatic Face and Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference*, pages 619–626, 2011.

[20] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, (3):21 – 45, 2006.

[21] H. Sloetjes, A. Russel, and A. Klassmann. Elan: a free and open-source multimedia annotation tool. In *INTERSPEECH*, pages 4015–4016. ISCA, 2007.

[22] M. Song, J. Bu, C. Chen, and N. Li. Audio-visual based emotion recognition - a new approach. In *CVPR (2)*, pages 1020–1025, 2004.

[23] K. M. Ting and I. H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:115–124, 1999.

[24] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'ericco, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3:69–87, April 2012. Issue 1.

[25] 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, Kloster Irsee, Germany*, LNCS, pages 188–199. Springer, 2008.

[26] J. Wagner, F. Lingenfelser, E. Andre, J. Kim, and T. Vogt. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218, 2011.

[27] 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 A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D. A. Shamma, M. Worring, and R. Zimmermann, editors, *ACM Multimedia*, pages 831–834. ACM, 2013.

[28] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *J. Sel. Topics Signal Processing*, 4(5):867–881, 2010.

[29] M. Wöllmer, M.llmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll. A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. *Neurocomput.*, 73(1-3):366–380, Dec. 2009.

[30] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, January 2009.

[31] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang. Audio-visual affective expression recognition through multistream fused hmm. *Trans. Multi.*, 10(4):570–577, June 2008.