J Multimodal User Interfaces (2014) 8:257–272
DOI 10.1007/s12193-014-0149-0

ORIGINAL PAPER

Combination of sequential class distributions from multiple channels using Markov fusion networks

Michael Glodek · Martin Schels · Friedhelm Schwenker · Günther Palm

Received: 27 March 2013 / Accepted: 6 February 2014 / Published online: 8 March 2014

Abstract The recognition of patterns in real-time scenarios has become an important trend in the field of multi-modal user interfaces in human computer interaction. Cognitive technical systems aim to improve the human computer interaction by means of recognizing the situative context, e.g. by activity recognition (Ahad et al. in IEEE, 1896–1901, 2008), or by estimating the affective state (Zeng et al., IEEE Trans Pattern Anal Mach Intell 31(1):39–58, 2009) of the human dialogue partner. Classifier systems developed for such applications must operate on multiple modalities and must integrate the available decisions over large time periods. We address this topic by introducing the Markov fusion network (MFN) which is a novel classifier combination approach, for the integration of multi-class and multi-modal decisions continuously over time. The MFN combines results while meeting real-time requirements, weighting decisions of the modalities dynamically, and dealing with sensor failures. The proposed MFN has been evaluated in two empirical studies: the recognition of objects involved in human activities, and the recognition of emotions where we successfully demonstrate its outstanding performance. Furthermore, we show how the MFN can be applied in a variety of different architectures and the several options to configure the model in order to meet the demands of a distinct problem.

M. Glodek (✉) · M. Schels · F. Schwenker · G. Palm
Institute of Neural Information Processing, Ulm, Germany
e-mail: michael.glodek@uni-ulm.de

M. Schels
e-mail: martin.schels@uni-ulm.de

F. Schwenker
e-mail: friedhelm.schwenker@uni-ulm.de

G. Palm
e-mail: guenther.palm@uni-ulm.de

Keywords Markov fusion network · Multi-modal data · Temporal multi-class problems · Robust classifier fusion · Multiple classifier systems

1 Introduction

Computational devices, and mobile devices in particular, have become an integral part of our everyday life. Recent trends show that these devices are increasingly equipped with sensors to perceive the environment of their users. By making use of algorithms from the field of pattern recognition, this data is used to improve personalized services and usabilities of applications and user interfaces. As a result, new research fields emerged in pattern recognition addressing the study of cognitive technical systems [7,53]. Companion systems are one of the most famous and recent representative of cognitive technical systems [54] and aim at making human computer interaction more efficient and comfortable by recognizing the situative context and emotional state of their users and to adapt to their habits. The functionality of Companion system relies on the development of suitable multi-modal information processing architectures which are able to meet the demanding requirements, such as: real-time processing, multinomial combination, tolerance to noisy and incomplete inputs.

The real-time requirement refers to the circumstance that the applied algorithms must provide a reasonable system response time which is comparable to human beings in order to sustain sensation of a natural interaction. However, besides the strong time constraint, continuous real-time processing bears also many advantages: adapt the classification models by incorporating the information gathered in the past, e.g. by using semi-supervised learning techniques [40,41,57]. Moreover, real-world scenarios often involve multi-class

problems, e.g. in case of emotion recognition [8,14] or action recognition [4,44]. A successful approach to address the challenging topic of multi-class problems is the application of hierarchical architectures which consistently provide class probability distributions to a next processing layer [17,33]. Studies have shown that a more robust recognition of complex patterns, e.g. complex user dispositions, can be derived by observing primitive behavioral cues [43,51]. In the same way "actions" can be regarded as basic building blocks to compose more sophisticated "activities" [19,21]. Scherer et al. [43] extended the concept and presented a generic architecture to combine multi-modal layered classifiers. The third requirement, the robustness against noisy inputs, is a problem in real-world scenarios in which the data is usually recorded under unrestrictive conditions, e.g. illumination, occlusion and background noise may change in the video and audio channels [22,24,25]. In addition, the system has to be able to handle sensor failures which may not only be caused by sensor malfunctions but also by the absence of a sensor signal, e.g. utterances perceived by the audio channel or facial expressions in the video channel. As a result, the extracted unimodal features and classifiers, often produce weak results or even failures. This issue can be addressed by taking multiple modalities into account. However, contradicting cues from different modalities can also further complicate the recognition conditions [34].

This article introduces a novel framework for the continuous combination of decisions from multiple classifiers. Typical multiple classifier systems (MCS) make use of standard combination methods such as: averaging, voting, decision templates, pseudo inverse and naive Bayes [26,29,47]. Thiel [50] showed that fusion algorithms using additional uncertainty information (e.g., fuzzy class membership [42], probabilistic quantities [49] or classifier confidences [38]) can outperform approaches that are based only on crisp class assignments. In real-time scenarios, these methods, which have been mainly proposed for static (non-temporal) classifier fusion, are often simply extended by additionally integrating decisions along the temporal axis over a time window [23,39]. However, the extension of classical approaches by utilizing time windows generally arises from the necessity to take respect of the real-time requirements and is often founded on the lack of alternative approaches. In literature, combiners being more sophisticated than the extension of standard approaches have not been studied comprehensively yet. In this context, Glodek et al. [18] proposed the application of Kalman filters, which are usually used in object tracking, for the combination of multi-modal classifier decisions in the scenario of emotion recognition by combining the multi-modal classifier decisions continuously over time. The study showed that the Kalman filter approach, can be successfully applied for fusion in a real-time scenario. Ramirez et al. [37] align the marginal probabilities of the unimodal classifiers outputs over time using a latent-dynamic conditional random field (LDCRF). This late fusion approach is realized by concatenating the sequential outputs which are then learned by the LDCRF in order to provide an improved mapping to the target labels. Pan et al. [35] proposed an approach to combine two hidden Markov models (HMM) which model the same event based on different modalities. This is realized by approximating the joint probability of the two observed sequences (each modeled by an independent HMM) using a maximum entropy principle. As a result, an additional probability needs to be computed, namely the probability of one observation sequence given the most likely sequence of hidden states of the other confronted HMM, which is the key part of the final probability distribution.

We propose a novel combination technique named Markov fusion network (MFN). In order to give an intuition of the key features, we consider the relation of the MFN to the application of Markov random fields (MRF) in image denoising [3]. Diebel and Thrun [10] proposed MRF to increase the resolution of a depth map by utilizing color gradients of an high-resolution image. An area of homogeneous colors smooths the corresponding area of the depth map utilizing the Markov assumption. However, image gradients allow stronger changes in depth map according to the proposition that objects are silhouetted against each other and therefore, relaxes the smoothing constraints. The algorithm takes advantage of the Markov assumption to smooth outliers and to reconstruct missing classifier decisions over time. The output stream estimated by the MFN is based on an arbitrary number of classifiers providing their decisions whenever possible. The MFN can be controlled externally by dynamically weighting the inputs or dynamically adjusting the smoothing parameters, likewise as proposed by Diebel and Thrun [10]. In order to address multi-class distributions, the MFN can additionally guarantee that the estimated output satisfies the axioms of probability theory.

The rest of this paper is structured as follows: In Sect. 2 we introduce and discuss the MFN algorithm. The two empirical studies conducted using the MFN are presented in Sect. 3. Each study is divided into sub-sections which are dealing with the data set, the MCS architecture and the results. Finally, the conclusion of the proposed fusion algorithm is drawn in Sect. 4.

2 Markov fusion network

The MFN is designed to combine time series of probability distributions of multiple individual classifiers. Hence, the input is given by $M \times T$ probability distributions over I classes where M denotes the number of classifiers and T the number of time steps. The probability distribution of classifier m at time step $t \in \mathcal{L}_m$ is given by $\mathbf{x}_{m,t} \in [0,1]^I$

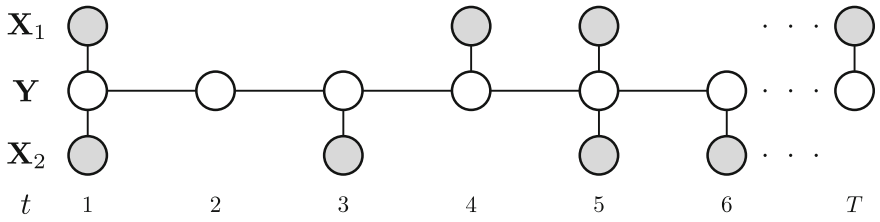


Fig. 1 Graphical model of the MFN. The estimates $\mathbf{y}_t$ are influenced by the the available decisions $\mathbf{X}_{m,t}$ of the source m and $t \in \mathcal{L}_m$ and adjacent estimates $\mathbf{y}_{t-1}$ and $\mathbf{y}_{t+1}$

where $m \in \{1, \ldots, M\}$, $\sum_{i=1}^{I} x_{m,i,t} = 1$ and $\mathcal{L}_m$ is the set of available probability distributions. The classifier predictions $\mathbf{X}_m \in [0,1]^{I \times T}$ are integrated by the MFN to a combined estimate $\mathbf{Y} \in [0,1]^{I \times T}$. Whenever class distributions are unavailable the corresponding node of the random variable and the connecting link is omitted.

The temporal fusion is accomplished by cliques realizing a Markov chain, which enforces the estimates being close in time to take similar values. Figure 1 depicts the graphical model of a MFN in which two sources of class distributions $\mathbf{X}_1$ and $\mathbf{X}_2$ are combined to an estimate $\mathbf{Y}$ for time $t = 1, \ldots, T$. The final estimate $\mathbf{y}_t$ is influenced by the available classifier decisions $\mathbf{x}_{mt}$ and the adjacent estimates $\mathbf{y}_{t-1}$ and $\mathbf{y}_{t+1}$. Therefore, temporal regions without any classifier decision are reconstructed by propagating the available information along the nodes.

We define the MFN by three potential functions, namely the *data potential* Ψ, the *smoothness potential* Φ and the *distribution potential* Ξ. The data potential Ψ is realized by a sum over the potentials Ψ_m which are enforcing the estimate of the time step t, i.e. $\mathbf{y}_t$, to be similar to the class distribution $\mathbf{x}_{m,t}$ of the mth classifier. The data potential is defined by

$$\Psi = \sum_{m=1}^{M} \Psi_m = \sum_{m=1}^{M} \sum_{i=1}^{I} \sum_{t \in \mathcal{L}_m} k_{m,t} (x_{m,i,t} - y_{i,t})^2,$$

where $\mathbf{K} \in \mathbb{R}_+^{M \times T}$ rates the reliability of the classifier m at time step t. The second potential Φ models the Markov chain and therefore, can enforce lateral (time discrete) smoothness. It is given by

$$\Phi = \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{\hat{t} \in N(t)} w_{\min(t,\hat{t})} (y_{i,t} - y_{i\hat{t}})^2,$$

where $\mathbf{w} \in \mathbb{R}_+^{T-1}$ weights the difference between two adjacent nodes and $N(t)$ returns the set of adjacent nodes, e.g. $N(t) := \{t-1, t+1\}$ in case both neighbors are available. In other words, the parameter $\mathbf{w}$ controls the strength of smoothing over time and can enforce neighboring predications to take similar values. Hence, the MFN is optimally suited whenever a continuous stream of decisions is required as an output. The third potential Ξ asserts that the resulting estimate is conform to the laws of probability theory and is given by

$$\Xi = u \cdot \sum_{t=1}^{T} \left(\left(1 - \sum_{i=1}^{I} y_{i,t} \right)^2 + \sum_{i=1}^{I} 1_{[0 > y_{i,t}]} \cdot y_{it}^2 \right)$$

where the parameter u weights the relevance of the potential and $1_{[0>y_{i,t}]}$ takes the value one in case $y_{i,t}$ is negative. The potential enforces the estimate to sum up to one for each time step and penalizes negative values.

The probability density function of the final estimate $\mathbf{Y}$ and the classifier predictions $\mathbf{X}_m$ is defined by

$$p(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_M) = \frac{1}{Z} \exp\Big(-\frac{1}{2} \big(\Psi + \Phi + \Xi \big) \Big)$$

where the partition function Z normalizes the probability to one. In order to determine the normalization constant Z, a considerable amount of computation has to be performed. Fortunately, only the mode of the density function is required such that the most likely estimate $\mathbf{Y}$ can be derived with only marginal computational effort.

To find the most likely sequence of predictions $\mathbf{Y}$ for a given set of parameters, we make use of an iterative gradient descent algorithm. In general, the gradient descent converges quickly to the desired solution in $O(T \cdot I \cdot M)$, confer Algorithm 1. However, in large regions of missing class distributions the lateral similarity might be propagated slowly. In these cases, the initial values of $\mathbf{Y}$ have to be chosen carefully. A suitable initialization is derived by taking the mean of the available probability masses for each time step and a linear interpolation in regions in which no distributions are available. As stopping criteria, we utilize a threshold based on the differences between two consecutive iteration steps. Since parameter learning requires an additional regularization term to handle the dependencies between partition parameters, the parameters of the presented study are derived by using grid search in the parameter space. However, this has not been addressed yet.

The MFN can be utilized *offline* by processing a complete recording of a fixed length at once. Alternatively, the MFN can be applied only to the most recent data utilizing a sliding window. In the following, we referred to the later approach as *online* processing.

2.1 MFN show-cases

We will demonstrate the effects of the parameters to the algorithm with the help of three artificially generated show cases.

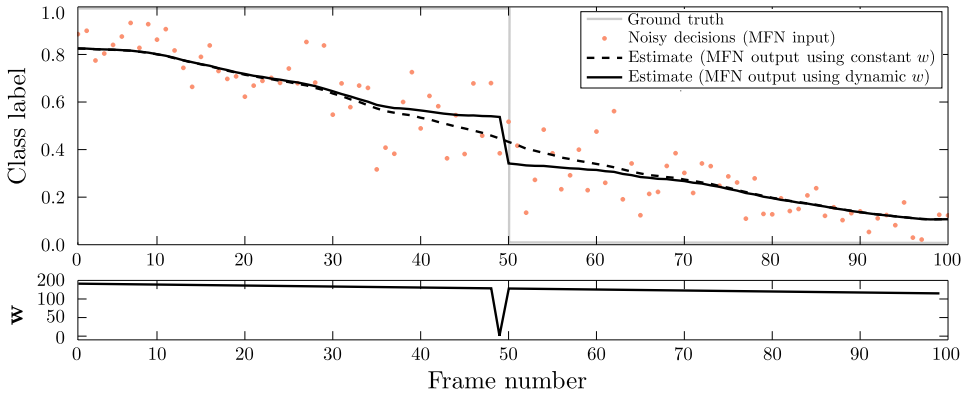


Fig. 2 Influence of the smoothness parameter **w** to the estimated decisions utilizing noisy classifier decisions as input to the MFN. For detailed description please refer to the text

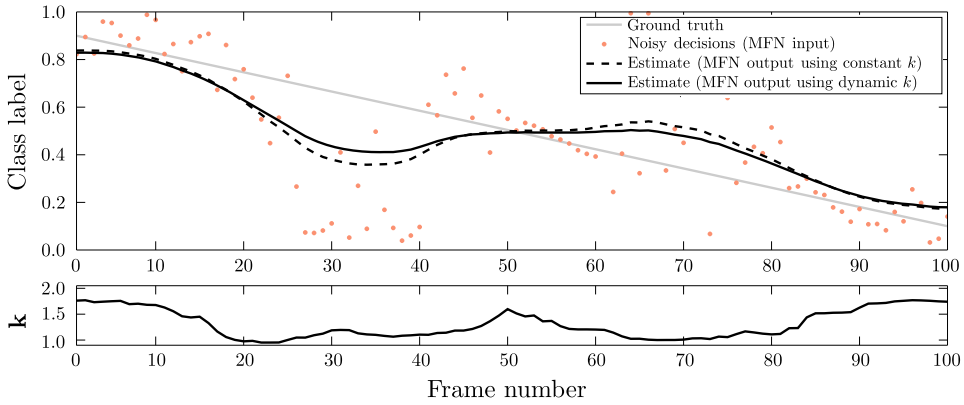


Fig. 3 Influence of the dynamic weighting of the parameter k_{tm} to the estimated decisions utilizing noisy classifier decisions as input to the MFN. For detailed description please refer to the text

Figure 2 depicts the influence of the parameter **w** which enforces the similarity over time.

The noisy decisions (light orange dots) represent a two-class distribution provided by a classifier (only one class of the probability distribution is shown). The assumed ground truth (gray line) shows a sudden change at the 50th frame and is only marginally covered by the input data. However, the recognition can be improved if additional information gives evidence that a class change is likely. The MFN results using constant values (dashed solid line) for the parameter **w** ($k_{t1} = 2$ for $t = 1 \ldots T$ and $w_t = 128$ for $t = 1 \ldots T - 1$) which offers a close fit to the orange dots. In case additional knowledge about the class change is available, it is possible to weaken the similarity by a single drop within the parameter **w**. The MFN results using dynamic parameters (solid black line) is based on the same parameter configuration with the exception of time step $t = 49$ in which we set $w_{49} = 0.5$ (the development of the parameter over time is shown in the lower part of the corresponding figure). Hence, the estimate makes a steep step downwards in order to minimize the energy function, which obviously is closer to the assumed ground truth function. The example showed the impact and possible applications of the smoothness parameter. The second example, shown in Fig. 3, exemplifies the control of the estimate using the data parameter **k**. Again the gray line represents the ground truth which is now given by a linear decay. The input distributions (orange dots) are locally distorted (again only one class of the probability distribution is depicted). Depending on the distance to the ground truth function, the level of distortion is increased. The dashed black line shows the estimate using $k_{t1} = 2.0$ for $t = 1 \ldots T$ and $w_t = 128$ for $t = 1 \ldots T - 1$ such that all classifier decisions effect the outcome with the same strength. In the given setting, it is self-evident to derive classifier confidence based on the standard deviation. For this purpose, we utilize a sliding window of ten frames to create a dynamic measure for weighting the input data via the data parameter **k**. The solid black line shows the estimate using the dynamic weighting of **k** as shown in the lower part of the figure. In regions of low variances the parameter k_{t1} takes values close to 2.0, whereas in regions of heavily scattered input, the parameter k_{t1} gets close to 1.0. The plot shows, that the output of the MFN using dynamic weighting is clearly less influenced by the outliers.

Figure 4 illustrates the third show case which points out the difference between online and offline processing.

The orange dots, which represent the artificial class distribution of the classifier, are closely distributed around the ground truth given by the gray curve. The dashed curve shows the estimate of the MFN processing the complete sequence at once, i.e. offline processing. In contrast, the solid line shows the MFN online processing which is obtained by calculat-

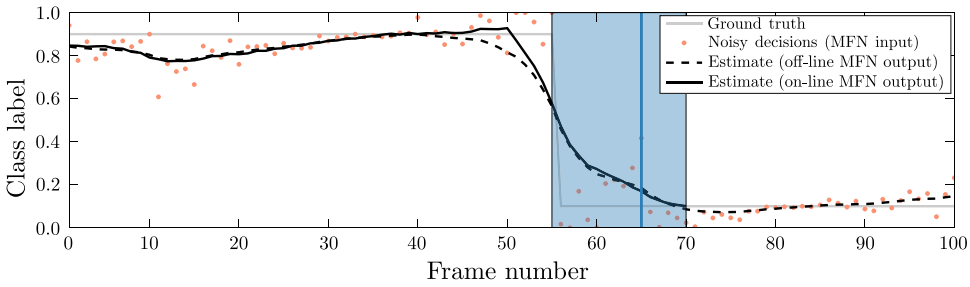


Fig. 4 Comparison of online and offline processing using the MFN. For detailed description please refer to the text

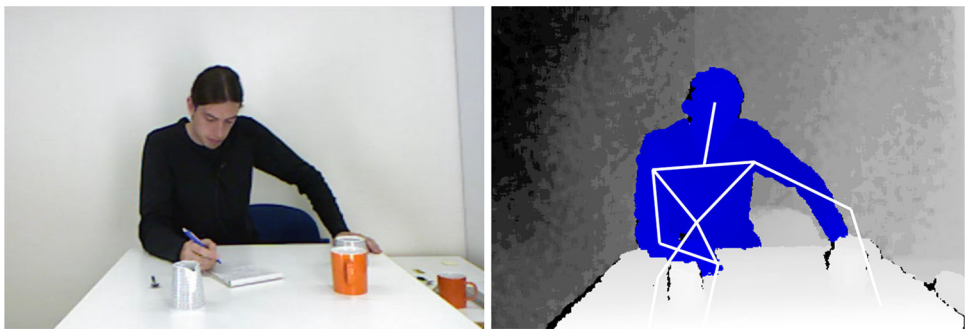


Fig. 5 Scene from the object recognition data set. The *left hand side* shows the RGB image, whereas the *right hand side* shows the depth map together with the tracked user (*colored in blue*) and the fitted skeleton (*white lines*)

ing the estimate using a sliding window (indicated by the filled light blue square expanding over 15 frames). For each frame, the MFN is fitted to the decisions of the corresponding window using the initialization of the preceding window. In case a slight delay is acceptable, it is advisable not to use the latest estimate as the final result, since an estimated value being somewhere placed in the middle of the window is less sensitive to outliers (in the figure we utilized the estimate at frame number ten within the window which is indicated by a vertical blue line).

The MFN offers a large set of advantageous properties: (i) it models the temporal relationship between probabilistic classifier predictions; (ii) it combines multiple classifier predictions from different modalities or views and is able to weight them dynamically; (iii) it handles missing classifier decisions, e.g. due to sensor failures; (vi) the estimate is consistent to the probability theory (v) it can easily be deployed in a real-time scenarios. The empirical studies presented in the next section will demonstrate these advantages and show possible application of the MFN.

3 Empirical studies

The MFN has been evaluated using two empirical studies, namely object recognition and emotion recognition. The first study is a multi-class scenario in which decisions from three feature views are combined to demonstrate the classical application of the MFN. The second study compares three different classifier fusion architectures in which the combination of the decisions is performed at different stages, i.e. early, mid-level and late fusion.

3.1 Object recognition task

The first experiment studies the combiner as part of a large human action recognition architecture, in which the performance of the system is enhanced by deriving additional information about the manipulated object.

3.1.1 Database description

The recording was performed in a desktop scenario using a retail Kinect™ camera[1]. The video stream is limited to 20 Hz and the audio signal is sampled at 16 kHz.

A sample frame of the data set is shown in Fig. 5. The RGB image on the left-hand side depicts a subject sitting in front of a desk on which objects are placed. The subject writes a note and therefore involving the manipulation of the pencil. On the right-hand side of the figure the corresponding depth map of the same scene is shown. The tracked subject is colorized in blue and overlay by the fitted skeleton.

In order to obtain a fixed-sized RGB sub-image of the currently used object, the location of the skeleton's right hand is utilized. As a result, the extracted image may contain other objects located close to the hand (for instance the milk can in Fig. 5). The recognition of objects is furthermore hindered by fast movements of the hand which result in blurred images or, by a minor delay in the skeleton fitting such that the sub-images depicts the hand only partially. Table 1 provides an overview of the short action sequences used for training and testing involving the objects: cup, milk can, paper, spoon,

[1] The Kinect™ camera is an input device developed by Microsoft®. http://www.xbox.com/en-US/Kinect (14/01/2014).

Table 1 Overview of the object recognition data set

Class name	Number of sequences	Avg. duration (s)
Cup	54	4.38 (0.92)
Milk can	53	4.64 (1.09)
Paper	57	5.49 (1.38)
Spoon	110	7.76 (2.16)
Pencil	60	10.94 (3.28)
Bare hand	237	1.05 (0.84)

Number of sequences and average duration (standard deviation in brackets) in seconds for each class

pencil and the bare hand, i.e the subject holds no object. The number of sequences per object range between 53 and 237. The class "bare hand" stands out because it is not related to any action but of short snippets and therefore, have a comparable large number of sequences. The average duration of the sequences, with exception of the class "bare hand", is given by the time required to perform the object-related action and ranges from around four up to eleven seconds.

Three types of features have been extracted from the subimage of the hand and the audio channel. From the image, we extracted an histogram of orientation gradients (HOG) [16] and an histogram of colors (HOC) [48]. The HOG uses a grid of 2×2 and 8 bins (each one covering 45°), which has been additionally normalized by the sum over all bins. The HOC is obtained in the hue-saturation-value (HSV) color space. A number of 10 bins are used for the hue dimension and 3 bins are used for the saturation dimensions (the dimension of value is omitted to realize light invariance). From the audio channel, we extracted the energy and mel frequency cepstral coefficients (MFCC) [24] using a 20 ms window with an offset of 10 ms. The energy is utilized for silence detection. The data set is partitioned into 10×10-fold cross-validation sets, meaning that the outer cross-validation, i.e. development and test set, are based on 10 folds and the cross-validation which is nested in these folds, i.e. the training and validation set, is as well comprised of 10 folds.

3.1.2 Architecture concepts

The proposed classifier architecture follows the concept of late classifier fusion [11,29] and is depicted in Fig. 6. The three features, i.e. MFCC, HOC and HOG, are extracted from the audio and video data and passed to corresponding classifiers operating on single frame level. The video-based object recognition relies on support vector machines (SVM) [9] which have additional probabilistic outputs according to Platt [36]. The training is performed in a one-vs-one manner resulting in 15 (6 choose 2) classifiers outputs. These outputs are projected to the final probability distribution using a pseudo-inverse trained on the set already used for training

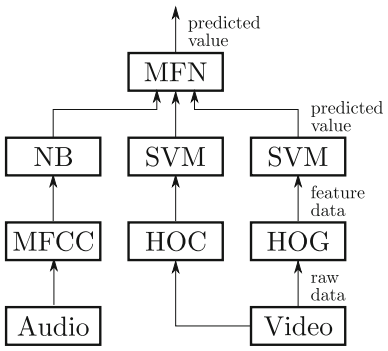


Fig. 6 Architecture design for the multi-modal multi-class recognition of objects. MFCC, HOC and HOG, are extracted from the raw audio and video data. These features are then fed into naive Bayesian classifiers (NB) and support vector machines (SVM) which render a probability distribution over the objects. These probability distributions are then combined by the MFN to obtain the final prediction

(in Fig. 6 the complete procedure is labeled by the abbreviation "SVM"). The audio-based object recognition utilizes a naive Bayes (NB) [27] classifier. Since most of the objects cause no sound when they are manipulated, only the manipulation using the spoon is recognized via the audio channel, i.e. when the spoon is used to stir the coffee in cup. In case no noise is present or the class spoon is not recognized the classifier returns no probability distribution. The three streams of probability distributions are combined by the MFN using offline processing.

3.1.3 Results

A small example shall provide an intuition of the inputs and outputs of the MFN algorithm. Figure 7 shows a test sequence in which the subject performs a series of actions involving multiple objects (the actual data is not part of the data set presented in the previous section). The actions are separated by pauses in which the hand is generally resting on the table without holding an object.

The upper three plots show the output of the video and audio classifiers, whereas the fourth plot shows the output of the MFN. The last plot shows the annotated ground truth. The plots depict frame-wise the share of probability given to the objects. Each color correspond to an object class (convey the legend of the figure for the color/object assignment). Consider for instance the MFN distribution (fourth plot) at frame 200. The highest probability mass is allocated on the class "pencil" with approximately 80 %. The second highest probability is set to the class "paper" with about 20 %. All other objects have only a marginal share of probability close to zero and therefore, these colors are not present in this frame. Frames in which no estimates are available are indicated by areas filled with diagonal lines. The two uppermost plots show the estimates of the SVM utilizing HOG and

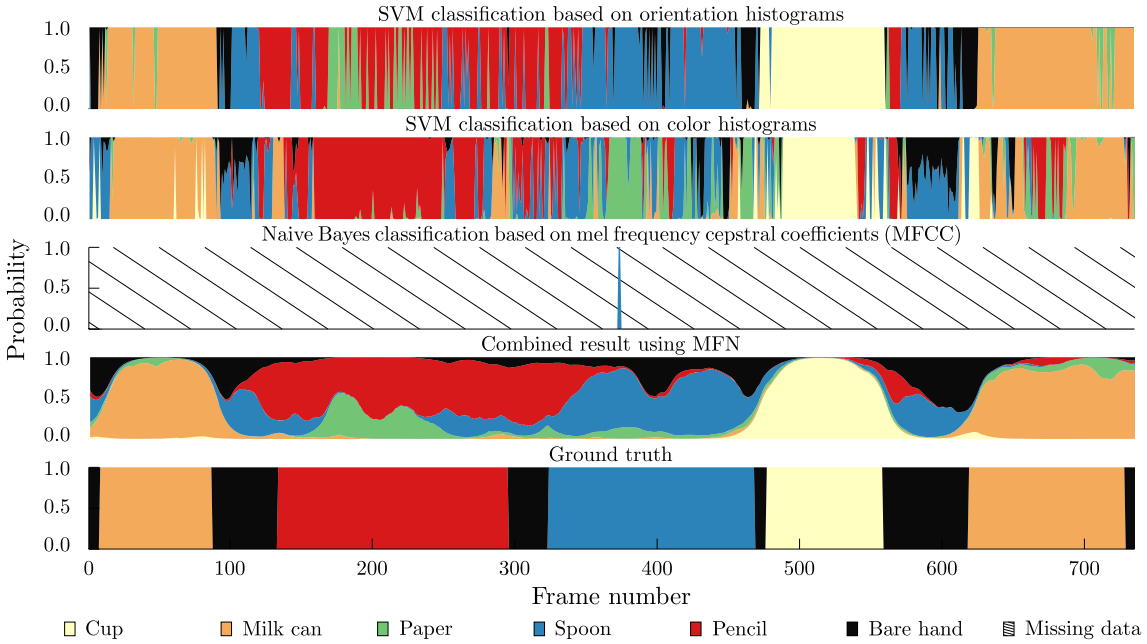


Fig. 7 Distribution over time resulting from all classifiers, the MFN and the labeled ground truth. For detailed description please refer to the text

HOC. Both classifiers return noisy probability distributions outputs which are often close to the ground truth. However, in distinct cases both classifiers tend to confuse the class membership, e.g. the class "paper". The third plot shows the prediction resulting from the audio channel using the naive Bayes classifier. Since in most frames the energy of the audio channel is lower then noise threshold, no estimates are available with the exception of the frames close to 385 in which the subject picks up the spoon and produces a sound by stirring the cup. The MFN, depicted in the fourth plot, combines the classifier decisions and simultaneously smooths them over time such that a large number of erroneous decisions are resolved. The MFN output is apparently close to the ground truth annotation shown in the last plot.

The quantitative evaluation is performed using the data set presented in the previous section. Table 2a shows the accuracies and the F_1 measures[2] of the classifiers and the MFN combination. The SVM utilizing the HOG features achieves an accuracy of 82.5 %, while the SVM using the HOC and the NB classifier based on the MFCC render only 65.7 and 53.5 %. The F_1 measures reveal the SVM classifiers provide a balanced recognition of all classes with the exception of the class "bare hand". As already mentioned, the hand generally rests on the table between the actions which often leads to a confusion to objects being close to the hand. The accuracy of the audio channel refers only to the class "spoon". The result of the MFN is listed in the last column and clearly outperforms the single classifiers by achieving an accuracy of 92.8 %. A similar finding can be drawn by examining the F_1 measures which are all ranging above 90 %; with the exception of the class "bare hand" which achieves only 80.7 %. However, the combination utilizing the MFN outperforms the input results in every aspect. For comparison, alternative fusion approaches have been evaluated: the sum and product of all modalities for every time step, i.e. point-wise fusion ($\perp$), and the sum and product of all modalities using a sliding window having a size of an eighth of the complete sequence, i.e. windowed fusion ($\sqcap$). The resulting windows have in average a length of 10.7 frames (10 frames median). The fusion is conducted by taking the sum, or respectively the product, of all available values. Subsequently, the quantities are renormalized to obtain a probability distribution over classes. The accuracies (listed in Table 2a) are better than the ones of the single modalities. However, the MFN fusion still performs better. Only the windowed sum fusion achieves an accuracy of 91.4 % which is close to the performance of the MFN fusion. However, in general the corresponding F_1 measures show a degraded performance, in particular the class "bare hand" reaches around 7 % less than the MFN.

Since this evaluation is based on a binarized rating of probability distribution, we extent the evaluation by an additional aspect. Table 2b depicts, the rankings achieved by the objects within the distributions. For instance, the accuracy concerning the first rank corresponds to the classical accuracy. The second rank denotes the share of frames in which the actual object was rated to be the second most likely object and so forth. The confusion of the SVM classifiers is considerably

[2] The F_1 measure is defined by $F_1 = 2\frac{P \cdot R}{P+R}$ where P is the precision and R the recall.

Table 2 Performance of the input classifiers, the point wise ($\perp$) and windowed ($\square$) product $\prod$ and average $\sum$ and the MFN

	HOG (video)	HOC (video)	MFCC (audio)	$\sum_{\perp}$ (fusion)	$\sum_{\square}$ (fusion)	$\prod_{\perp}$ (fusion)	$\prod_{\square}$ (fusion)	MFN (fusion)
(a)								
↑Acc.	82.5 (2.9)	65.7 (2.7)	53.5 (10.1)	82.2 (2.5)	92.4 (1.5)	84.2 (3.0)	84.0 (4.7)	92.8 (2.5)
↑F_1(Cup)	92.2 (3.0)	69.5 (2.5)	n/a	87.9 (2.9)	95.9 (2.4)	92.9 (2.9)	66.9 (6.9)	95.8 (2.9)
↑F_1(Milk can)	83.2 (5.3)	60.7 (4.4)	n/a	81.5 (4.8)	92.4 (3.7)	85.2 (5.1)	85.6 (5.1)	94.7 (2.7)
↑F_1 (Paper)	84.6 (4.5)	57.2 (3.7)	n/a	83.5 (4.4)	91.7 (1.9)	85.3 (4.3)	88.5 (2.8)	93.4 (3.8)
↑F_1(Spoon)	83.7 (2.5)	65.3 (4.0)	69.2 (8.5)	82.7 (2.4)	93.9 (1.7)	85.6 (3.1)	90.1 (4.0)	94.2 (2.5)
↑F_1(Pencil)	82.4 (5.1)	76.7 (4.3)	n/a	85.0 (4.4)	95.7 (2.3)	86.3 (5.1)	87.2 (9.1)	93.7 (4.5)
↑F_1(Bare hand)	64.6 (5.5)	47.1 (2.9)	n/a	64.6 (4.7)	73.1 (4.9)	63.6 (5.6)	70.0 (5.6)	80.7 (6.2)
(b)								
↑Acc. (rank 1)	82.5 (2.9)	65.7 (2.7)	53.5 (10.1)	82.1 (2.5)	92.4 (1.5)	84.2 (3.0)	92.4 (1.5)	92.8 (2.5)
↓Acc. (rank 2)	10.7 (1.1)	17.2 (1.3)	46.5 (10.1)	9.9 (1.1)	4.1 (0.8)	9.0 (1.2)	4.1 (0.8)	5.5 (1.1)
↓Acc. (rank 3)	3.7 (1.0)	7.2 (0.9)	n/a	4.0 (0.6)	2.1 (0.5)	3.8 (0.7)	2.1 (0.5)	1.4 (1.3)
↓Acc. (rank 4)	1.6 (0.8)	4.3 (0.5)	n/a	2.0 (0.5)	0.8 (0.4)	1.5 (0.8)	0.8 (0.4)	0.3 (0.3)
↓Acc. (rank 5)	0.8 (0.5)	3.3 (0.7)	n/a	1.2 (0.4)	0.3 (0.3)	0.7 (0.4)	0.3 (0.3)	0.0 (0.1)
↓Acc. (rank 6)	0.8 (1.1)	2.3 (0.7)	n/a	0.8 (0.4)	0.2 (0.3)	0.9 (1.2)	0.2 (0.3)	0.0 (0.1)

All values in percent with standard deviation in brackets. Arrows indicate how to rate the values of the corresponding measures. (a) Accuracies and per-class F_1 measures of each classifier and the MFN. (b) Average rank accuracy. The rank denotes the ordered positions achieved in the probability distribution

smaller than initially expected by regarding the accuracies of Table 2a alone, since the first two ranks together acquire already around 93.2 and 82.9 %. Again the decisions based on the audio channel are difficult to evaluate since, they are only present in small portion of frames and assign their probabilities for only one class. The MFN combination issues the correct class among the first two highest probabilities with a chance of 98.3 %.

3.1.4 Results summary

The presented setting demonstrated the application of the MFN in a multimodal recognition task in which a single modality can provide decisions only when they are available. The fusion over modalities and over time clearly improved the performance of the unimodal classifiers. When compared to extensions of standard fusion approaches, i.e. point-wise or windowed fusion using a sum or a product, the MFN shows as well better results. Only the windowed fusion summing the available probabilities achieves a comparable but slightly decreased accuracy. With respect to ranking, the results reveal that the correct class is very likely to be within the first ranks of the MFN, when compared to the windowed fusion using the sum.

3.2 Emotion recognition task

Emotion recognition is an emerging field which has received much attention over the last years. First findings in emotion recognition are made mostly based on acted recordings of emotion [13,34]. The utilization of acted emotions has many advantages such as they (1) inherit a valid ground truth; (2) are recorded under controlled conditions; and (3) contain prominent features to distinguish the emotions. In recent time, the research focus has changed to more challenging non-acted recordings [8,28,43,45]. Typically, non-acted recordings have a larger variety of expressed emotions, which are generally more difficult to interpret, and exceptional events, being not related to the target classes, are more likely to occur, e.g. scratching the nose or turning away from the camera. Especially the annotation of ground truth required for training appears to be very challenging since on the one hand the occurring emotions are strongly related to the targeted application and on the other hand the emotions are often ambiguous [34]. In summary, the automated recognition of non-acted emotions bears many more challenges compared to acted recordings.

Many promising approaches have been proposed, addressing the problem of emotion recognition. For instance, an increasing number of classifier systems have been developed addressing multi-modal architectures, due to the finding that emotions are inherently expressed via multiple channels, e.g. facial and verbal expressions [56]. These approaches promise to be more robust and above all less affected by events which are not directly related to the emotional state. Furthermore, latest classifier systems tend to incorporate an end-to-end handling of uncertainty, to counteract the adverse conditions of the uncertainties associated with the recordings and the ground truth [20,42]. Furthermore, recent advances in classifier systems make use of context-sensitive information, e.g. transitions between emotional states or conversational turns [43,52,55].

The current study aims at examining three different classifier architectures for emotion recognition making use of audio and video-based classifiers having probabilistic outputs. The architectures utilize the MFN to combine the classifier outputs with additional context information, i.e. the conversational turns, to form an intermediate or final fusion result. Furthermore, the performance of the classifier outputs are accessed using a confidence measure such that weak classifier decisions can be rejected. An additional confidence value over time is also used to weight the classifier outputs with the help of the data potential. The rejection of decisions having a low confidence aims at improving the fusion by elevating the quality of input decisions. The experiment shows that the MFN can be utilized in a broad range of possible architectures and be extended without any effort to integrate additional information.

3.2.1 Database description

The Audio/Visual Emotion Challenge (AVEC) 2011 was introduced in the context of the ACII 2011 workshop and is composed of 95 audio-visual recordings of 13 subjects who are interacting with affectively colored artificial agents [31,46]. The data set is labeled in four affective dimensions: "arousal", "expectancy", "power" and "valence" [15]. For each dimension, the annotations of the raters are averaged, resulting in a real value for each time step. Subsequently, the labels are binarized using a threshold equal to the grand mean of each dimension such that the classes and their negations are equally balanced. Every recording was annotated by two to eight raters. Along with the sensor data and annotation, a word-by-word transcription of the spoken language was provided which partitions the dialog into conversational turns. The AVEC 2011 is divided into three sub-challenges: an audio challenge on word-level, a video challenge on frame-level and an audiovisual also on video frame-level[3] [31,46].

For the challenge, the data set has been divided into a training set, a development set and a test set (of which the annotation has not been published yet). In order to derive a meaningful statistic for the present study, the recordings have been repartitioned to 4×4 cross-validation sets where each fold is limited to a distinct set of persons, such that no training on test subjects is performed. Although the challenge provided a large range of precomputed features, we decided to extract a new customized set of features for both modalities.

Three different features are extracted from the raw audio signal: fundamental frequency, mel frequency cepstral coefficient and *perceptual linear predictive* [23,24].

[3] A comprehensive description and the data can be found at http://sspnet.eu/avec2011/ (14/01/2014).

The video features are extracted using the modules "Basic Emotions 4.4.3", "FACS 4.4", "Unilaterals" and "Smile Detector" of the computer expression recognition toolbox (CERT) [30]. The output is the concatenation to an overall 36-dimensional vector per frame. CERT requires a successful detection of the face in order to track these features. However, because of the unrestricted recording settings, e.g. subjects may turn away or leave the visual range of the camera, the detection fails in about 8 % of the frames.

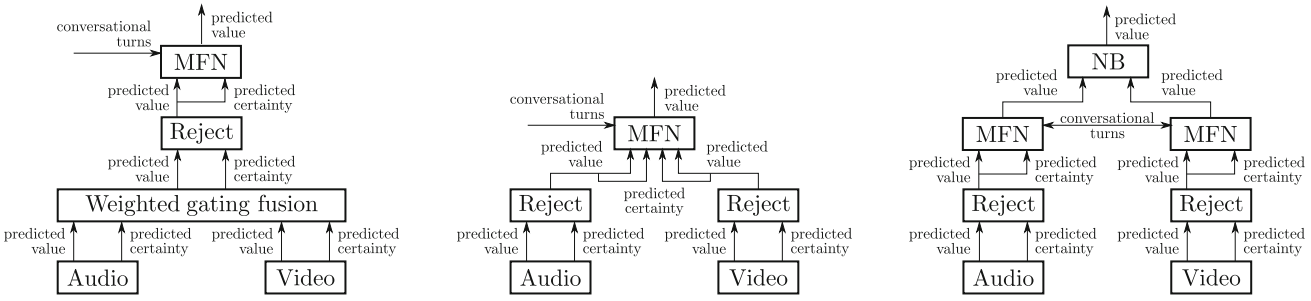
3.2.2 Architecture concepts

The proposed architecture requires robust classifiers to be trained for each channel which provide an additional confidence measure per class predication. We achieved the robustness by utilizing bagging [5] and made use of simple classifier functions which have proven to be well-suited for the recognition of the task at hand. Bagging is performed by training classifiers on different sub-sets of the data. The final predication is then obtained by averaging the individual classifier outputs [5]. We derive the classifier confidence based on the standard deviation of the individual classifier outputs. The stronger the consensus of the classifiers, the higher the confidence is set. The video classification uses five naive Bayes classifiers for bagging [5], The audio classifiers operate on word-level (the information about the word location is provided by the transcript) using a fixed-length representation of the features. A corresponding transformation is obtained with the help of HMM according to [2]. Bagging is then applied using five random forests [6]. The classification results are used for all architectures without modifications.

The architectures utilized aim at comparing different design concepts making use of the MFN. In particular, we focus on the following aspects:

1. Early, mid-level and late fusion architectures
2. A real-world scenario with sensor failures
3. Classifiers using the reject option
4. Temporal integration of context information (conversational turns)
5. Additional weighting of predictions by certainty measures
6. Online and offline implementations of architectures.

The processing of sequential data using MFN offers a large variety of possible classifier architectures. We follow the systematic categorization proposed by Dietrich [12] and decompose the classifier system into three components, namely the F-step (fusion of decisions), the R-step (rejection of uncertain decisions) and the T-step (temporal integration of the decisions). By varying the F-step, we obtain three architectures FRT, RFT and RTF as shown in Fig. 8 which we will name early, mid-level and late-fusion from the view point of clas-



(a) Early fusion (FRT) architecture. **(b)** Mid-level fusion (RFT) architecture. **(c)** Late fusion (RTF) architecture.

Fig. 8 Three fusion architectures FRT, RFT and RTF evaluated. For detailed description please refer to the text

sifier outputs. Each modality, i.e. audio and video, provides a prediction of the target in form of a class membership accompanied by a measure of how certain the prediction is (confidence). The certainty of the decision is utilized to decide which prediction to reject or not. The rejection is performed based on a rejection rate which defines the ratio of decisions to be discarded.

The fusion of the modalities of every architecture is performed differently. In case of the early fusion (FRT), the decisions and the confidences are combined before the rejection and temporal fusion is conducted, as shown in Fig. 8a. For each frame, the available decisions and confidences are averaged. In case no decisions are provided by any classifier, the final frame will be marked as having no decision available. Due to the early and rather rudimentary fusion, no outstanding performance can be anticipated. Figure 8b illustrates the mid-level (RFT) architecture which first rejects audio and video decisions separately and then performs simultaneously a multi-modal and temporal fusion using the MFN. The late fusion (RTF) architecture, shown in Fig. 8c, implements the multi-modal fusion after the temporal integration using a naive Bayes classifier trained on the set which has also be utilized to determine the optimal MFN parameters (no parameters optimization took place for the naive Bayes classifiers). The additional training step bears the danger of over-fitting. However, since a very large amount of data is available and the distribution of class decisions is already close to the final result, the risk is rather low. The improvement will most likely balance the class distribution of the decisions which might have been shifted due to the rejection and reconstruction of the decisions.

All approaches make use of additional information: (1) temporal integration of context information; (2) additional weighting of predictions using certainty measures. Furthermore, all architectures are evaluated utilizing the offline and online processing. The temporal integration of context information is realized by decreasing the weighting w_t each time a speaker-turn occurs, i.e. every-time the speaker changes during the conversation. The additional weighting of predictions is performed by measuring the standard deviation σ of the classifier outputs within a window of two seconds shifted over the sequence. The confidence measure is derived by evaluating $1 - \sigma$ multiplied by a ratio of how many decisions of the window are available (in case the window contains no classifier outputs, the confidence is set to zero). The obtained confidence is then multiplied to the corresponding data potential parameter **k**. The online processing is based on a window of five seconds in which the MFN combines the input data. The resulting estimate of the forth second is then stored as the final output. To speed-up the convergence, the MFN utilizes the outputs of the last estimate as the initial input shifted according to the elapsed time.

3.2.3 Unimodal results

Table 3 shows the performance of the base classifiers for audio and video. Accuracies and the F_1 measures of the classifiers without rejection are provided in Table 3a. The accuracies of all label dimensions are close to 60 %. With the exception of "valence", the audio classifiers appear to have a slightly better recognition performance. The best performing class, according to the accuracies and F_1 measures, is "arousal". "Expectancy", "power" and "valence" of the audio classifier show a slightly unbalanced distribution of classes, as to be confirmed by the F_1 measures. Table 3b, c list the classifier performances for rejecting 10 and 90 % of the decisions (decisions having the lowest confidence are rejected first). Generally, the audio and video classifiers are able to improve their recognition rates slightly. However, the improvement when reject decisions with low confidence directly results in a sparse streams of classifier predications which have to be reconstructed by the MFN. The F_1 measures remain rather stable indicating that the class distribution remains almost unchanged.

In the next step, we evaluate the performance of the MFN applied to single modalities for online and offline processing.

The results in Table 4 have been evaluated on all frames of the test set recordings. Best performing parameters and rejec-

Table 3 Performance of word-wise and frame-wise results of the audio and video modality for different rejection rates

	Arousal	Expect.	Power	Valence
(a) No rejection				
Audio				
↑Acc.	61.8 (3.6)	58.9 (6.3)	57.5 (9.4)	57.5 (7.9)
↑F_1	65.8 (3.8)	16.4 (7.1)	69.6 (9.3)	70.1 (6.8)
↑$\overline{F}_1$	56.7 (3.4)	72.6 (5.2)	24.7 (6.6)	24.9 (8.4)
Video				
↑Acc.	57.0 (4.3)	54.7 (4.0)	55.7 (2.8)	59.9 (7.4)
↑F_1	60.9 (5.1)	49.6 (9.4)	57.4 (11.3)	67.1 (11.5)
↑$\overline{F}_1$	51.3 (9.3)	56.6 (10.7)	48.7 (12.2)	43.5 (7.1)
(b) 10 % rejection rate				
Audio				
↑Acc.	62.0 (3.5)	58.7 (6.3)	57.5 (9.4)	57.5 (7.6)
↑F_1	66.1 (3.8)	15.8 (7.3)	69.6 (9.3)	70.1 (6.6)
↑$\overline{F}_1$	56.9 (3.3)	72.5 (5.2)	24.5 (6.2)	24.4 (8.8)
Video				
↑Acc.	57.7 (4.3)	55.0 (3.9)	55.7 (2.8)	60.4 (7.8)
↑F_1	61.5 (5.2)	50.3 (9.5)	57.2 (11.3)	67.6 (11.9)
↑$\overline{F}_1$	51.8 (9.3)	56.6 (10.6)	49.1 (12.2)	43.5 (7.5)
(c) 90 % rejection rate				
Audio				
↑Acc.	62.7 (3.1)	59.1 (6.6)	59.1 (9.9)	57.5 (7.6)
↑F_1	67.3 (2.8)	15.9 (8.8)	71.1 (9.4)	70.1 (6.6)
↑$\overline{F}_1$	56.4 (4.3)	72.7 (5.6)	24.2 (5.3)	24.4 (8.8)
Video				
↑Acc.	59.5 (4.3)	56.9 (1.6)	54.4 (4.1)	67.8 (10.4)
↑F_1	65.0 (5.1)	56.0 (13.5)	54.9 (10.3)	75.6 (11.8)
↑$\overline{F}_1$	50.6 (9.3)	52.5 (14.0)	50.7 (10.0)	41.0 (12.7)

All values in percent with standard deviation in brackets. Arrows indicate how to rate the values of the corresponding measures

Table 4 Performance of the reconstructed unimodal streams using the MFN

	Arousal	Expect.	Power	Valence
(a) Audio				
Online				
↑Acc.	71.7 (6.4)	56.2 (6.7)	58.3 (9.8)	60.2 (10.6)
↑F_1	74.7 (6.2)	5.6 (3.5)	69.6 (9.4)	72.9 (8.3)
↑$\overline{F}_1$	67.3 (7.6)	71.3 (5.5)	30.7 (7.0)	21.6 (13.3)
A. rej. (%)	10	50	90	50
Offline				
↑Acc.	68.7 (6.3)	56.4 (6.3)	55.5 (10.4)	59.0 (10.5)
↑F_1	72.6 (5.6)	9.5 (3.6)	69.4 (9.7)	72.8 (8.1)
↑$\overline{F}_1$	63.1 (8.6)	71.1 (5.5)	10.9 (12.4)	11.9 (13.8)
A. rej. (%)	0	90	0	50
(b) Video				
Online				
↑Acc.	63.4 (1.9)	58.7 (3.9)	59.0 (4.9)	65.8 (9.5)
↑F_1	66.2 (5.3)	55.2 (11.6)	60.7 (13.9)	73.2 (11.7)
↑$\overline{F}_1$	57.0 (11.8)	57.9 (14.0)	48.4 (16.7)	45.9 (7.0)
V. rej. (%)	50	90	0	90
Offline				
↑Acc.	62.5 (4.0)	57.9 (4.8)	58.5 (5.0)	64.4 (9.7)
↑F_1	65.0 (4.8)	51.5 (12.7)	59.9 (14.2)	72.0 (11.3)
↑$\overline{F}_1$	56.6 (14.2)	58.3 (14.7)	48.1 (16.8)	45.7 (9.2)
V. rej. (%)	50	10	0	90

All values in percent with standard deviation in brackets. Arrows indicate how to rate the values of the corresponding measures

tion rates are determined based on the training set (rejection rates listed in the Table 4 while the parameters of the MFN are given in Table 7a in the Appendix). Compared to the results of the raw input shown in Table 3 the reconstruction renders similar or improved performance. For instance, the classification of "arousal" shows a clearly increased performance with an accuracy of up to 71.1 % in the audio channel and 63.4 % in the video channel. The classification of "expectancy", "power" and "valence" remains fairly stable, although decisions are now available for all frames. The F_1 measures of the audio channel for "expectancy", "power" and "valence" indicate that the classifier output has become increasingly unbalanced which can be related to the sparseness and uneven distribution of the input decisions. Compared to the audio channel, the video channel is clearly more stable concerning the F_1 measures. An interpretation of the selected rejection rates cannot be drawn easily because of the close interplay with MFN parameters and the distribution of the remaining decisions. For instance, the more decisions are rejected, the smaller is the impact of the MFN parameter of the data potential. Furthermore, we see that the offline processing clearly outperforms the online processing which can be attributed to the small window used. The offline procedures utilize strong smoothing parameters such that a large amount of information of the past and of the future is integrated. The window used in online processing is limited to small segments of five seconds and no information of future events is available.

3.2.4 Multimodal results

The results of the FRT, RFT and RTF architectures, as depicted in Fig. 8, are listed in Table 5a–c. The Table 5a shows the performance of the first classifier architecture. As anticipated, only marginal improvement can be achieved, compared to the unimodal fusion. While the classification of "arousal" and "valence" perform worse, "expectation" and "power" performs slightly better as seen in Table 4. However, due to the multiple modalities which are incorporated

Table 5 Performance of the three multi-modal classifier architecture

	Arousal	Expect.	Power	Valence
(a) Architecture FRT (Fig. 8a)				
Online				
↑Acc.	66.8 (4.4)	60.3 (3.9)	58.4 (3.7)	62.9 (8.7)
↑F_1	70.5 (3.5)	47.9 (10.6)	64.8 (8.4)	73.3 (8.5)
↑$\overline{F}_1$	61.5 (8.0)	66.7 (7.1)	44.5 (11.8)	33.3 (12.5)
Offline				
↑Acc.	66.8 (4.4)	60.0 (3.9)	58.6 (8.0)	62.2 (9.2)
↑F_1	70.7 (3.7)	48.5 (11.3)	67.7 (10.1)	72.7 (8.8)
↑$\overline{F}_1$	61.1 (8.2)	65.8 (7.5)	38.3 (4.4)	32.3 (12.4)
A. rej. (%)	10	90	90	90
V. rej. (%)	90	90	90	90
(b) Architecture RFT (Fig. 8b)				
Online				
↑Acc.	68.8 (5.2)	62.1 (2.9)	61.0 (6.0)	64.3 (9.0)
↑F_1	72.5 (4.6)	46.5 (12.7)	67.2 (8.7)	72.9 (10.5)
↑$\overline{F}_1$	63.4 (8.6)	69.6 (5.2)	45.7 (14.9)	40.3 (10.3)
Offline				
↑Acc.	68.2 (4.6)	60.9 (4.1)	59.9 (6.2)	62.6 (9.2)
↑F_1	72.2 (4.3)	42.0 (11.9)	64.4 (10.6)	72.3 (9.6)
↑$\overline{F}_1$	62.4 (7.5)	69.6 (5.7)	46.4 (16.7)	35.2 (12.3)
A. rej. (%)	10	50	90	90
V. rej. (%)	50	90	0	90
(c) Architecture RTF (Fig. 8c)				
Online				
↑Acc.	68.9 (8.2)	59.2 (5.2)	54.6 (3.3)	64.1 (9.1)
↑F_1	66.7 (10.0)	41.2 (8.1)	57.2 (1.8)	68.1 (14.7)
↑$\overline{F}_1$	70.3 (7.9)	68.7 (4.3)	50.7 (9.4)	54.0 (7.3)
Offline				
↑Acc.	65.8 (5.6)	56.8 (8.7)	51.4 (6.7)	63.4 (10.2)
↑F_1	68.7 (2.5)	28.9 (17.4)	63.4 (4.0)	69.3 (14.9)
↑$\overline{F}_1$	61.0 (11.8)	67.8 (8.7)	25.9 (18.2)	46.9 (11.2)
A. rej. (%)	10	50	90	50
V. rej. (%)	50	90	0	90

All values in percent with standard deviation in brackets. Arrows indicate how to rate the values of the corresponding measures

into a unified decisions, the classifier system can provide decisions based on recent information even if a modality becomes inoperative. Both, the online and offline procedures render similar accuracies and F_1 measures. Table 5b shows the results of the second architecture (Fig. 8b) which implements the temporal and multi-modal fusion at once. Compared to the previous architecture, the accuracies have been clearly improved and outperform the unimodal results of the classification of "expectancy" and "power". The classification of "arousal" and "valence" has a lower recognition performance than the uni-lateral fusion which can be related to the weighted mixture of both modalities. The information which has been discarded by the FRT architecture by decomposing the multi-modal and temporal fusion can be retained by making use of the simultaneous fusion in the MFN. The results of the last architecture RTF are shown in Table 5c and reveal a similar performance for the dimensions "arousal" and "valence", i.e. 68.9 and 64.1.2 %, compared to the RFT architecture and a decreased performance of 59.2 and 54.6 % for the dimensions "expectancy" and "power". The F_1 measures indicate that the outputs are in general more balanced then the outputs of the RFT architecture and the unimodal fusion. However, the circumstance that less information in each channel is available over all time steps during the fusion results in a decrease of the accuracies.

Table 6 Comparison of the results to the AVEC 2011 baseline and winner of the challenge

↑*Avg. Acc.*	Audio	Video	Audio-visual
Baseline (%)	51.95	47.50	53.73
Winner (%)	57.65 (UCL)	60.71 (USC)	n/a
MFN online (%)	61.60	61.73	64.05 (RFT)

The challenge winner has been assessed by averaging all the four accuracies of the affective dimensions. The audio sub-challenge was won by the University College London (UCL) and the video sub-challenge was won by the University of Southern California (USC). The audio-visual challenge was not evaluated due to lack of participants. The table shows the unweighted accuracies for the baseline, the winner and the online MFN. The MFN online results have been raised on a repartitioned set using the available development and training set. All values in percent. Arrows indicate how to rate the values of the corresponding measures

3.2.5 Comparison to AVEC 2011 results

Table 6 shows the obtained results in comparison with the official challenge baseline and winner. The challenge has been rated by averaging the accuracies of all four affective dimensions. The baseline of the audio sub-challenge was set to an accuracy of 51.95 %. The winner of this sub-challenge was Meng and Bianchi-Berthouze [32] who achieved an average unweighted accuracy 53.65 %. The MFN utilizing the online fusion approach outperforms the winner having an accuracy of 61.60 %. However, it is important to emphasis that all MFN results are based on a repartitioned set of the available development on training set, as explained in detail in Sect. 3.2.1. The video sub-challenge has been won by Ramirez et al. [37] with an average unweighted accuracy of 60.71 % whereas the baseline was 47.50 %. The online MFN achieves an accuracy of 61.73 %. Concerning the audio-visual challenge no winner was chosen because of an insufficient number of participants. However, the baseline was set to 53.73 %. In this scenario, the online MFN based on the RFT architecture is able to obtain an accuracy of 64.05 %.

3.2.6 Results summary

The fusion can improve the recognition of the dimensions "expectancy" and "power", whereas the unimodal fusion of "arousal" and "valence" perform better (similar results have been obtained in previous related studies [23]). We address this issue to the distribution of decisions provided by the modalities: While the decisions of the modalities with respect to "arousal" and "valence" apparently contradict each other in regions of temporal closeness, the decisions of the modalities with respect to "expectancy" and "power" complement each other. Furthermore, the balancing effect of the naive Bayes classifier shows that further improvements can be achieved by developing an algorithm to adapt the potential parameters of the MFN. We believe that an algorithm having optimal constant weights for $\mathbf{K}$ and $\mathbf{w}$ can be realized in the near future in case suitable assumptions are made, e.g. the ratio of available and unavailable decisions remains stable for each modality. A dynamical adaption of weights could also help to resolve contradicting decisions. However, assessing decisions in such a way might only be possible using an heuristic algorithm and additional information.

4 Conclusions

In the field of pattern recognition multi-modal classifier systems attracted researchers in various real-time scenarios, e.g. activity or emotion recognition [1,43,56].

We propose a novel temporal fusion algorithm, namely the Markov fusion network (MFN), which is especially designed for such kind of applications. The model exhibits three major properties (1) weighting the individual decisions; (2) enforcing temporal similarity; and (3) ensuring that the axioms of probability distributions hold. Due to its variable structure, model parameters may change dynamically, furthermore the MFN mitigate sensor failures.

The model has been evaluated in two empirical settings: object and emotion recognition. In the first study, we presented the combination of three classifier outputs in the scenario of a multi-class problem, where one classifier contributed to the fusion by providing decisions for only one class. The results of the MFN clearly outperformed each single classifier. The second study addresses the field of emotion recognition, here the flexibility of the MFN architecture is presented. Three different architectures have been evaluated in detail. It has been shown that the MFN being used to combine the multi-modal decisions together with temporal information simultaneously performed best among the proposed architectures.

The successful application of the MFN demonstrated in this work, show the capabilities and the flexibility of the proposed model. Future work will aim at exploring new areas of use and performing a comprehensive analysis of possible MFN configurations. Furthermore, we envisaged the development of a learning algorithm for the MFN to directly estimate suitable parameters for the potentials.

Acknowledgments This paper is based on work done within the Transregional Collaborative Research Centre SFB/TRR 62 Companion-Technology for Cognitive Technical Systems funded by the German Research Foundation (DFG).

Appendix

This section contains the gradient decent algorithm to determine the MFN estimate (Algorithm 1), and the parameters used within the second experiment (Table 7).

ALGORITHM 1: Gradient descent algorithm to find the most likely estimate $\mathbf{Y}$ of the MFN.

Input: Initial estimate $\mathbf{Y} \in [0,1]^{I\times T}$, classifier predictions $\mathbf{X} \in [0,1]^{M\times I\times T}$ and model parameters $\mathbf{w} \in \mathbb{R}_+^{T-1}$, $\mathbf{K} \in \mathbb{R}_+^{M\times T}$ and $u \in \mathbb{R}_+$.
Output: Energy E and the gradient $\frac{\partial \mathbf{Y}}{\partial y_{it}}$.

$E_{\text{data}} = 0$; $E_{\text{smooth}} = 0$; $E_{\text{dist}} = 0$;
for *t=1* **to** T **do**
  $s = 0$
  for *i=1* **to** I **do**
    $s = s + y_{it}$
  for *i=1* **to** I **do**
    $g = 0$
    for *m=1* **to** M **do**
      if isavailable(x_{mit}) **then**
        $E_{\text{data}} = E_{\text{data}} + k_{mt} \cdot \left(y_{it} - x_{mit}\right)^2$
        $g = g + 2 \cdot k_{mt} \cdot \left(y_{it} - x_{mit}\right)$
    if $t == 0$ **then**
      $E_{\text{smooth}} = E_{\text{smooth}} + w_0 \cdot \left(y_{i0} - y_{i1}\right)^2$
      $g = g + 4 \cdot w_0 \cdot \left(y_{i0} - y_{i1}\right)$
    else if $t < T$ - *1* **then**
      $E_{\text{smooth}} = E_{\text{smooth}} + w_{t-1} \cdot \left(y_{it} - y_{it-1}\right)^2 + w_t \cdot \left(y_{it} - y_{it+1}\right)^2$
      $g = g + 4 \cdot w_{t-1} \cdot \left(y_{it} - y_{it-1}\right) + 4 \cdot w_t \cdot \left(y_{it} - y_{it+1}\right)$
    else if $t == T$ - *1* **then**
      $E_{\text{smooth}} = E_{\text{smooth}} + w_{T-2} \cdot \left(y_{iT-1} - y_{it-2}\right)^2$
      $g = g + 4 \cdot w_{T-2} \cdot \left(y_{iT-1} - y_{it-2}\right)$
    $E_{\text{dist}} = E_{\text{dist}} + u \cdot \left((1-s)^2 + (1_{[0<y_{it}]} \cdot y_{it})^2\right)$
    $g = g + u \cdot 2 \cdot \left((s-1) + 1_{[0<y_{it}]} \cdot y_{it}\right)$
  $\frac{\partial \mathbf{Y}}{\partial y_{it}} = g$
$E = E_{\text{data}} + E_{\text{smooth}} + E_{\text{dist}}$

Table 7 Emotion recognition parameter setting of the MFN

	A.	E.	P.	V.
(a) Unimodal (Table 4)				
Audio				
w_u	400	96	400	256
w_t	32	96	32	32
k_{audio}	0.1	0.2	0.1	0.2
Video				
w_u	400	128	400	96
w_t	32	4	32	96
k_{video}	0.1	1.0	1.0	1.0
(b) Multi-modal (Table 5)				
FRT				
w^u	400	400	400	400
w^t	32	32	32	32
k	0.5	0.5	0.5	0.5
λ_{audio}	1.0	0.1	0.1	1.0
λ_{video}	0.1	1.0	1.0	1.0
RFT				
w^u	400	128	400	96
w^t	32	4	32	96
k_{audio}	0.1	1.0	1.0	1.0
k_{video}	0.1	1.0	1.0	1.0
RTF				
w^u_{audio}	400	96	128	96
w^t_{audio}	32	96	4	4
k_{audio}	0.5	0.3	0.1	0.1
w^u_{video}	400	128	400	96
w^t_{video}	32	4	32	96
k_{video}	0.05	0.3	0.05	0.3

References

1. Ahad MAR, Tan J, Kim H, Ishikawa S (2008) Human activity recognition: various paradigms. In: Proceedings of the international conference on control, automation and systems (ICCAS). IEEE, pp 1896–1901. doi:10.1109/ICCAS.2008.4694407
2. Bicego M, Murino V, Figueiredo M (2003) Similarity-based clustering of sequences using hidden Markov models. In: Proceedings of the international conference on machine learning and data mining (MLDM), Lecture Notes in Computer Science (LNCS), vol 2734. Springer, Berlin, pp 95–104. doi:10.1007/3-540-45065-3-8
3. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
4. Brand M, Oliver N, Pentland A (1997) Coupled hidden Markov models for complex action recognition. In: Proceedings of the international IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 994–999. doi:10.1109/CVPR.1997.609450
5. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. doi:10.1007/BF00058655
6. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. doi:10.1023/A:1010933404324
7. Buss M, Beetz M, Wollherr D (2007) CoTeSys—cognition for technical systems. In: Proceedings of the COE workshop on human adaptive mechatronics (HAM)
8. Castellano G, Leite I, Pereira A, Martinho C, Paiva A, McOwan PW (2010) Affect recognition for interactive companions: challenges and design in real world scenarios. J Multimodal User Interfaces 3(1–2):89–98. doi:10.1007/s12193-009-0033-5
9. Christiani N, Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge University Press, Cambridge
10. Diebel J, Thrun S (2006) An application of Markov random fields to range sensing. In: Proceedings of advances in neural information processing systems (NIPS), vol 18. MIT Press, Cambridge, pp 291–298
11. Dietrich C, Palm G, Riede K, Schwenker F (2004) Classification of bioacoustic time series based on the combination of global and local decisions. Pattern Recognit 37(12):2293–2305. doi:10.1016/j.patcog.2004.04.004
12. Dietrich CR (2004) Temporal sensorfusion for the classification of bioacoustic time. Ph.D. thesis, Institut of Neural Information Processing, University of Ulm, Ulm, Germany
13. Douglas-Cowie E, Campbell N, Cowie R, Roach P (2003) Emotional speech: towards a new generation of databases. Speech Commun 40(1–2):33–60. doi:10.1016/S0167-6393(02)00070-5
14. Ekman P (1992) An argument for basic emotions. Cognit Emot 6(3–4):169–200
15. Fontaine J, Scherer K, Roesch E, Ellsworth P (2007) The world of emotions is not two-dimensional. Psychol Sci 18(12):1050
16. Freeman W, Roth M (1995) Orientation histograms for hand gesture recognition. Tech. Rep. TR94-03, Mitsubishi Electrical Research Laboratories. Originally published at the International Workshop on Automatic Face and Gesture Recognition
17. Glodek M, Bigalke L, Schels M, Schwenker F (2011) Incorporating uncertainty in a layered HMM architecture for human activity recognition. In: Proceedings of the joint workshop on human gesture and behavior understanding (J-HGBU). ACM, pp 33–34. doi:10.1145/2072572.2072584
18. Glodek M, Reuter S, Schels M, Dietmayer K, Schwenker F (2013) Kalman filter based classifier fusion for affective state recognition. In: Zhou ZH, Roli F, Kittler J (eds) Multiple classifier systems (MCS), Lecture Notes in Computer Science (LNCS), vol 7872. Springer, Berlin, pp 85–94. doi:10.1007/978-3-642-38067-9_8
19. Glodek M, Schels M, Palm G, Schwenker F (2012) Multiple classifier combination using reject options and Markov fusion networks. In: Proceedings of the international ACM conference on multimodal interaction (ICMI). ACM, pp 465–472. doi:10.1145/2388676.2388778
20. Glodek M, Scherer S, Schwenker F (2011) Conditioned hidden Markov model fusion for multimodal classification. In: Proceedings of the annual conference of the international speech communication association (Interspeech). ISCA, pp 2269–2272
21. Glodek M, Schwenker F, Palm G (2012) Detecting actions by integrating sequential symbolic and sub-symbolic information in human activity recognition. In: Perner P (ed) Proceedings of the international conference on machine learning and data mining (MLDM), Lecture Notes in Computer Science (LNCS), vol 7376. Springer, Berlin. pp 394–404. doi:10.1007/978-3-642-31537-4_31
22. Glodek M, Trentin E, Schwenker F, Palm G (2013) Hidden Markov models with graph densities for action recognition. In: Proceedings of the international joint conference on neural networks (IJCNN). IEEE, pp 964–969
23. Glodek M, Tschechne S, Layher G, Schels M, Brosch T, Scherer S, Kächele M, Schmidt M, Neumann H, Palm G, Schwenker F (2011) Multiple classifier systems for the classification of audio-visual emotional states. In: D'Mello S, Graesser A, Schuller B, Martin JC (eds) Affective computing and intelligent interaction,

Lecture Notes in Computer Science (LNCS), vol 6975. Springer, Berlin, pp 359–368. doi:10.1007/978-3-642-24571-8_47
24. Huang X, Acero A, Hon H (2001) Spoken language processing: a Guide to Theory. Prentice Hall, Algorithm and System Development
25. Kim M, Kumar S, Pavlovic V, Rowley H (2008) Face tracking and recognition with visual constraints in real-world videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1–8. doi:10.1109/CVPR.2008.4587572
26. Kittler J, Hatef M, Duin RP, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239. doi:10.1109/34.667881
27. Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. The MIT Press, Cambridge
28. Krell G, Glodek M, Panning A, Siegert I, Michaelis B, Wendemuth A, Schwenker F (2012) Fusion of fragmentary classifier decisions for affective state recognition. In: Schwenker F, Scherer S, Morency LP (eds) Multimodal pattern recognition of social signals in human-computer-interaction, Lecture Notes in Computer Science (LNCS), vol 7742. Springer, Berlin, pp 116–130. doi:10.1007/978-3-642-37081-6_13
29. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley , New York. doi:10.1002/0471660264
30. Littlewort G, Whitehill J, Wu T, Fasel I, Frank M, Movellan J, Bartlett M (2011) The computer expression recognition toolbox (CERT). In: Proceedings of the international conference IEEE on automatic face gesture recognition and workshops (FG). IEEE, pp 298–305. doi:10.1109/FG.2011.5771414
31. McKeown G, Valstar M, Cowie R, Pantic M (2010) The SEMAINE corpus of emotionally coloured character interactions. In: Proceedings of the international conference on multimedia and expo (ICME). IEEE, pp 1079–1084. doi:10.1109/ICME.2010.5583006
32. Meng H, Bianchi-Berthouze N (2011) Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In: D'Mello S, Graesser A, Schuller B, Martin JC(eds) Proceedings of the international conference on affective computing and intelligent interaction (ACII), Lecture Notes in Computer Science (LNCS), vol 6975. Springer, pp 378–387. doi:10.1007/978-3-642-24571-8_49
33. Oliver N, Garg A, Horvitz E (2004) Layered representations for learning and inferring office activity from multiple sensory channels. Comput Vis Image Underst 96(2):163–180. doi:10.1016/j.cviu.2004.02.004. (Special issue: Event Detection in video)
34. Palm G, Glodek M (2013) Towards emotion recognition in human computer interaction. In: Esposito A, Squartini S, Palm G (eds) Neural nets and surroundings, smart innovation, systems and technologies, vol 19. Springer, pp 323–336. doi:10.1007/978-3-642-35467-0_32
35. Pan H, Levinson S, Huang T, Liang ZP (2004) A fused hidden Markov model with application to bimodal speech processing. IEEE Trans Signal Process 52(3):573–581. doi:10.1109/TSP.2003.822353
36. Platt J (2000) Probabilistic outputs for SV machines, chap. 5. Neural Information Processing Series. MIT Press, Cambridge, pp 61–74
37. Ramirez GA, Baltrušaitis T, Morency LP (2011) Modeling latent discriminative dynamic of multi-dimensional affective signals. In: D'Mello S, Graesser A, Schuller B, Martin JC (eds) Proceedings of the international conference on affective computing and intelligent interaction (ACII), Lecture Notes in Computer Science (LNCS), vol 6975. Springer, pp 396–406. doi:10.1007/978-3-642-24571-8_51
38. Schels M, Glodek M, Meudt S, Scherer S, Schmidt M, Layher G, Tschechne S, Brosch T, Hrabal D, Walter S, Palm G, Neumann H, Traue H, Schwenker F (2013) Multi-modal classifier-fusion for the recognition of emotions. In: Coverbal synchrony in Human-Machine Interaction. CRC Press, pp 73–97
39. Schels M, Glodek M, Meudt S, Schmidt M, Hrabal D, Böck R, Walter S, Schwenker F (2012) Multi-modal classifier-fusion for the classification of emotional states in WOZ scenarios. In: Ji YG (ed) Advances in affective and pleasurable design, vol 22 in Advances in Human Factors and Ergonomics Series. CRC Press, pp 644–653. doi:10.1201/b12525-78
40. Schels M, Kächele M, Glodek M, Hrabal D, Walter S, Schwenker F (2013) Using unlabeled data to improve classification of emotional states in human computer interaction. J Multimodal User Interfaces 1–12. doi:10.1007/s12193-013-0133-0 (Special Issue: From Multimodal Analysis to Real-Time Interactions with Virtual Agents)
41. Schels M, Kächele M, Hrabal D, Walter S, Traue H, Schwenker F (2012) Classification of emotional states in a Woz scenario exploiting labeled and unlabeled bio-physiological data. In: Schwenker F, Trentin E (eds) Proceedings of the international conference on partially supervised learning (PSL), Lecture Notes in Computer Science (LNCS), vol 7081. Springer, pp 138–147. doi:10.1007/978-3-642-28258-4_15
42. Schels M, Scherer S, Glodek M, Kestler H, Palm G, Schwenker F (2013) On the discovery of events in EEG data utilizing information fusion. Comput Stat 28(1):5–18. doi:10.1007/s00180-011-0292-y
43. Scherer S, Glodek M, Layher G, Schels M, Schmidt M, Brosch T, Tschechne S, Schwenker F, Neumann H, Palm G (2012) A generic framework for the inference of user states in human computer interaction: How patterns of low level behavioral cues support complex user states in HCI. J Multimodal User Interfaces 6(3–4):117–141. doi:10.1007/s12193-012-0093-9
44. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: Proceedings of the international conference on pattern recognition (ICPR), vol 3. IEEE, pp 32–36
45. Schuller B, Seppi D, Batliner A, Maier A, Steidl S (2007) Towards more reality in the recognition of emotional speech. In: Proceedings of the international IEEE conference on acoustics, speech and signal processing (ICASSP), vol 4. IEEE, pp 941–944. doi:10.1109/ICASSP.2007.367226
46. Schuller B, Valstar M, Eyben F, McKeown G, Cowie R, Pantic M (2011) AVEC 2011—the first international audio visual emotion challenges. In: D'Mello S, Graesser A, Schuller B, Martin JC (eds) Proceedings of the international conference on affective computing and intelligent interaction (ACII), Lecture Notes in Computer Science (LNCS), vol 6975. Springer, pp 415–424. doi:10.1007/978-3-642-24571-8_53 (Part II)
47. Schwenker F, Dietrich CR, Thiel C, Palm G (2006) Learning of decision fusion mappings for pattern recognition. J Artif Intell Mach Learn 17–21 (Special issue: Multiple Classifier Systems)
48. Swain M, Ballard D (1991) Color indexing. Int J Comput Vis 7(1):11–32
49. Szczot M, Löhlein O, Palm G (2012) Dempster-Shafer fusion of context sources for pedestrian recognition. In: Denoeux T, Masson MH (eds) Belief functions: theory and applications, advances in intelligent and soft computing, vol 164. Springer, pp 319–326
50. Thiel C (2010) Multiple classifier systems incorporating uncertainty. Verlag Dr, Hut
51. Vinciarelli A, Pantic M, Bourlard H, Pentland A (2008) Social signal processing: State-of-the-art and future perspectives of an emerging domain. In: Proceedings of the international ACM conference on multimedia (MM). ACM, pp 1061–1070. doi:10.1145/1459359.1459573
52. Vlasenko B, Schuller B, Wendemuth A, Rigoll G (2007) Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In: Paiva AC, Prada R, Picard RW (eds) Proceedings of the internation conference on affective computing and intelligent interaction (ACII), Lecture Notes in Computer

Science (LNCS), vol 4738. Springer, pp 139–147. doi:10.1007/978-3-540-74889-2_13
53. Wahlster W (2003) SmartKom: symmetric multimodality in an adaptive and reusable dialogue shell. In: Krahl R, Günther D (eds) Proceedings of the status conference "Human Computer Interaction". DLR, pp 47–62
54. Wendemuth A, Biundo S (2012) A companion technology for cognitive technical systems. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller VC (eds) Cognitive behavioural systems, Lecture Notes in Computer Science (LNCS), vol 7403. Springer, pp 89–103. doi:10.1007/978-3-642-34584-5_7
55. Wöllmer M, Metallinou A, Eyben F, Schuller B, Narayanan S (2010) Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In: Proceedings of the annual conference of the international speech communication association (ISCA), interspeech, pp 2362–2365
56. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58. doi:10.1109/TPAMI.2008.52
57. Zhu X (2005) Semi-supervised learning literature survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison