Neurocomputing 161 (2015) 17–37

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Fusion paradigms in cognitive technical systems for human–computer interaction

Michael Glodek [a], Frank Honold [b], Thomas Geier [c], Gerald Krell [d], Florian Nothdurft [e], Stephan Reuter [f], Felix Schüssel [b], Thilo Hörnle [c], Klaus Dietmayer [f], Wolfgang Minker [e], Susanne Biundo [c], Michael Weber [b], Günther Palm [a], Friedhelm Schwenker [a]

[a] *Institute of Neural Information Processing, University of Ulm, Germany*
[b] *Institute of Media Informatics, University of Ulm, Germany*
[c] *Institute of Artificial Intelligence, University of Ulm, Germany*
[d] *Institute of Information Technology and Communications, Otto-von-Guericke University Magdeburg, Germany*
[e] *Institute of Information Technology, University of Ulm, Germany*
[f] *Institute of Measurement, Control, and Microtechnology, University of Ulm, Germany*

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form
7 January 2015
Accepted 9 January 2015
Available online 27 February 2015

Keywords:
Cognitive technical systems
Scene understanding
Human–computer interaction

ABSTRACT

Recent trends in human–computer interaction (HCI) show a development towards cognitive technical systems (CTS) to provide natural and efficient operating principles. To do so, a CTS has to rely on data from multiple sensors which must be processed and combined by fusion algorithms. Furthermore, additional sources of knowledge have to be integrated, to put the observations made into the correct context. Research in this field often focuses on optimizing the performance of the individual algorithms, rather than reflecting the requirements of CTS. This paper presents the information fusion principles in CTS architectures we developed for Companion Technologies. Combination of information generally goes along with the level of abstractness, time granularity and robustness, such that large CTS architectures must perform fusion gradually on different levels — starting from sensor-based recognitions to highly abstract logical inferences. In our CTS application we sectioned information fusion approaches into three categories: perception-level fusion, knowledge-based fusion and application-level fusion. For each category, we introduce examples of characteristic algorithms. In addition, we provide a detailed protocol on the implementation performed in order to study the interplay of the developed algorithms.

1. Introduction

Modern computer systems are designed to improve on efficiency and user experience by dynamically adapting to situations, incorporating additional knowledge and enhancing the interaction. These features are realized by enabling the computer to perceive its environment, to extract relevant information and to compare this to previously acquired data. In the literature cognitive technical systems (CTS) are known as Companion Systems. State-of-the-art systems available on the market claim to provide these kinds of features, however they are still far below their possible potential, mostly due to the demanding information processing required [1,2].

Perception in a CTS can be divided into three major categories which are virtually omnipresent in any given human–computer interaction (HCI) setting: (1) the implicit user input [3] (e.g. emotion or disposition [4]); (2) the explicit user input (e.g. multimodal instructions by gesture and speech); and (3) the recognition of the user's environment as well as the context of use [5] (e.g. activities, state and manipulation of objects nearby). It must be emphasized that necessary perceptions usually strongly depend on the application at hand. Therefore, it is always important to identify the relevant and application-specific perceptions in a first step. In case of emotions, useful classes are not necessarily the most obvious ones, e.g. happiness or anger. In fact, to improve an interaction it is more beneficial to focus on negative user

E-mail addresses: michael.glodek@uni-ulm.de (M. Glodek), frank.honold@uni-ulm.de (F. Honold), thomas.geier@uni-ulm.de (T. Geier), gerald.krell@ovgu.de (G. Krell), florian.nothdurft@uni-ulm.de (F. Nothdurft), stephan.reuter@uni-ulm.de (S. Reuter), felix.schuessel@uni-ulm.de (F. Schüssel), thilo.hoernle@uni-ulm.de (T. Hörnle), klaus.dietmayer@uni-ulm.de (K. Dietmayer), wolfgang.minker@uni-ulm.de (W. Minker), susanne.biundo@uni-ulm.de (S. Biundo), michael.weber@uni-ulm.de (M. Weber), guenther.palm@uni-ulm.de (G. Palm), friedhelm.schwenker@uni-ulm.de (F. Schwenker).

http://dx.doi.org/10.1016/j.neucom.2015.01.076

dispositions being directly related to the system, like boredom or stress [2].

1.1. Fusion categories in cognitive technical systems

In CTS, the problems often arise from the endeavor to develop architectures which implement multiple requirements simultaneously. The perception of classes, especially of the second and the third category, is encumbered by the open world scenario in which unusual events may occur and classes often have a wide range of variability. This perception problem can be addressed by enriching the recognition approach with additional domain knowledge. However, what appears to be a straightforward solution entails many open research questions. The most important one is How to realize a seamless integration of symbolic and sub-symbolic information, also with respect of how to exchange information in both directions, i.e. from sensory to high-level representations and back. Furthermore, it is not sufficient that these algorithms perform well on convenient pre-segmented datasets but have to provide good results in real-time in ubiquitous applications. In turn, more requirements arise: how to (1) compensate sensor failures; (2) draw information from the temporal dimension; (3) take uncertainty into account; and (4) deal with the open world setting. However, the problems mentioned so far only address the perceptive periphery. Another central issue represents the combination of the uncertain perceptions with symbolic domain knowledge. The combination is crucial to enhance the recognition results and to bring them into the correct context. The integration of domain knowledge is also inevitable, since the recording of datasets covering all possible observations is usually infeasible.

In addition, the inferred classes are typically more abstract, which is helpful to create a truly relevant user history and which in turn is necessary to carry out a reasonable interaction. Indeed, the concept of combining explicit user inputs with knowledge about the ongoing and past behavioral patterns of the user bears another challenge. In a first step, an abstract input representation has to be derived from inputs of multiple modalities which then has to be combined with available knowledge, the dialog management and the application. In other words, an algorithm has to mediate between the user's input, acquired knowledge (possibly afflicted with uncertainty), goal priorities, and the interface provided by the application.

In the last instance, the CTS's fission component has to reason about how to provide a situation and input dependent output based on an abstract representation of the core system. The presented work addresses these challenges and shows ways how it is possible to solve them by taking advantages from information fusion methods.

Shifting the view from outer requirements to the characteristics of information, it becomes evident that the processing can be grouped into different stages. Fig. 1 exemplifies the stages and how the information is condensed by fusion. Algorithms close to the sensory usually recognize patterns which are directly observable in the scene, e.g. the presence or the identification and attributes of a person [6,7]. By adding spatial and temporal context, more meaningful classes can be derived, e.g. recognizing activities or the mood of a person [8,4]. In the next layer, relationship of entities can be taken into account, e.g. persons with respect to each other or connection between a person and objects [9]. Again, a large history of observations can allow the discovery of more complex attitudes and salient events. Ultimately, this kind of high-level information is of relevance for the application and interaction. We regard the decisions based on the high-level information as the last step of fusion. The requirements and characteristics motivate the partitioning of the fusion algorithms and architectures into these categories: *perception-level fusion*, *knowledge-based fusion* and *application-level fusion*. The new taxonomy will be used throughout the paper as an aid to orientation and explained in greater detail in the corresponding sections.

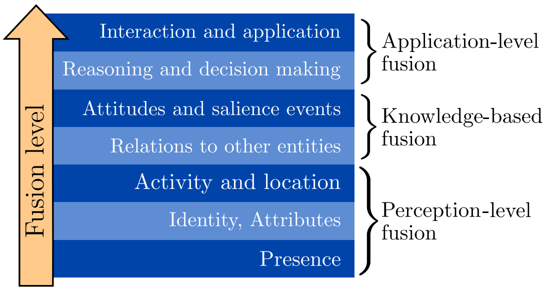


Fig. 1. *Information fusion in CTS grouped into three layers: perception-level fusion, knowledge-based fusion and application-level fusion.* The higher the layer of fusion, the more abstract the derived and processed knowledge. The procedure is usually accompanied with an increase of the temporal granularity and the variability of the occurrences covered.

1.2. Architecture of cognitive technical systems

An alternative view on CTS is to take a closer look at its architecture design. The systematical decomposition of a CTS architecture is depicted in Fig. 2. The schema shows the exchange of information between the user and the system, where the red arrows represent the input and the blue arrows the output of the system. Basically, the system itself is organized in two basic blocks: (1) peripheral block consisting of the user interface and perception component; and (2) an inner block consisting of a knowledge model and associated components such as planning, ontologies and the application and dialog management itself.

The red input arrow in the lower right, which is leading into the perception component of the first basic block, represents the recognizers perceiving the environment and the intrinsic user state. The multimodal inputs, e.g. video cameras or microphones, are mapped to classes by the perception component. In case one class is recognized by multiple modalities, the perception-level fusion combines them to a single output. The perception component is connected to the knowledge model, not only to derive more sophisticated information but also to enhance the perception by back-propagating beliefs. This is achieved by the knowledge-base fusion, represented by the lower bi-directional arrow. On the upper right red and blue arrows represent the input and the output of the system, respectively. The commands, which are combined and interpreted by the user interface, are, if necessary, forwarded to the inner basic block which then adapts the knowledge model and planning accordingly. The bi-directional arrows in the system show that in the ideal case information is exchanged in both directions in a seamless manner. Therefore, the bi-directional arrows represent not only the fusion of information but also fission

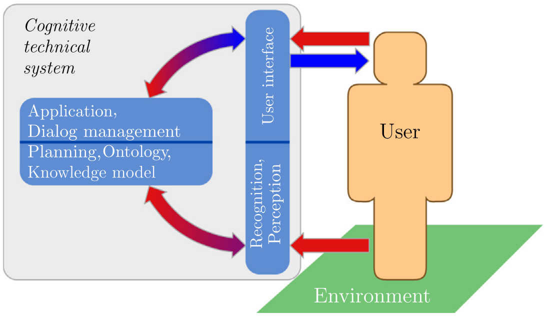

Cognitive technical system
User interface
Application, Dialog management
Planning, Ontology, Knowledge model
Recognition, Perception
User
Environment

Fig. 2. *Architecture design of a CTS.* The CTS perceives two kinds of input, namely the implicit input (lower arrow) and the explicit input (upper arrow). Within the CTS, the information needs to be processed gradually with respect to the temporal granularity, the level of abstractness and uncertainty in order to allow a robust extraction. (For interpretation of the references to color in the text, the reader is referred to the web version of this paper.)

[10], in which high-level knowledge is transferred back in order to enhance performance of the components close to the periphery. The combination of information on this level is referred to as application-level fusion.

1.3. Related work

The development of CTS has a long history in computer science [11] and many different approaches have been proposed so far [12–15]. Newell [16], one of the cognitive pioneers, defined multiple criteria which an architecture with human-like cognition has to satisfy to be functional [17]. These criteria show that the goals in creating a CTS are ambitious and that still no feasible approach has been found, which is capable to come close to human cognition [17,15]. In general, there are two main directions of development, namely the cognitivist paradigm which is based on symbolic information processing, and the emergent paradigm which embraces connectionist systems, dynamical systems and enactive systems [18]. Most approaches put focus on one of these two paradigms. However, considerable effort has as well been put on studying hybrid models, which combine emergent and cognitivist systems [18–20] and what is also followed in the work presented in this paper. According to our opinion it is important to differentiate between the overall concept of an architecture and its basic components, since the components might be as well applicable in other architectures. Hence, we aim at providing a clear and comprehensive overview by first discussing cognitive architecture designs and then later on an overview of state-of-the-art components.

One of the best-known cognitive architectures is SOAR which systematically decomposes functional properties to components, e.g. working memory or production memory. SOAR is largely rule-based, and therefore a clear representative of the cognitivist paradigm [13].

The ACT theory started to evolve shortly after and states that complex cognition arises from an interaction of procedural and declarative knowledge. It generalizes former approaches and provides a powerful, flexible framework to set up cognitive systems [21] which is still being under development and constantly extended [22].

Fink et al. [23,24] proposed a distributed system for integrated speech and image understanding. The system first evaluates the multimodal input to extract basic knowledge, e.g. object and word recognition and perceptual grouping. At a higher level of the system, the scene and speech is analyzed to generate a robot control and synthesize language output. A unique feature is given by the fact that the speech recognition takes results of the understanding component into account in order to generate its predictions, which means that a bi-directional exchange of information is realized.

The SmartKom project [25,26] aims at providing an adaptive and modular framework for multimodal interaction. The central component of the architecture is the interaction management which receives the fused multimodal sensor input in order to produce the multimodal output of the system. The interaction management is furthermore bi-directionally interconnected with the application which is accessed via a generic interface. The final instantiations of the generic architecture strongly depend on the interaction and application requirements [26] and show a large variety. However, the main flow of information from perception to the output rendering can be regarded as being uni-directional.

Burghart et al. [27] proposed a three-layered cognitive architecture for humanoid robots where the higher layers correlate with higher complexity and understanding. The modular design allows fast reaction times to external events while having an explicit integration of goals in the higher planning layers. Furthermore, the architecture is characterized by a strong bi-directional interconnection of adjacent modules.

The CoTeSys project, which investigates cognition for technical systems such as vehicles, robots and factories, proposed an alternative architecture [11] in which a planning and control module is used as the central component. It receives information from the perception, learning/reasoning and knowledge/models and sends the control to the linked actuators. Additional edges from the perception to the learning/reasoning and from the learning/reasoning to the knowledge/models enable the system to adapt to the environment and to the human. The project advocates strong interconnection between these modules to achieve the synergies and to realize a sophisticated cognitive system.

1.4. Outline

CTS depend on algorithms being able to combine data from heterogeneous sources. The combination of such kind of data allows the acquisition of new high-level knowledge which will be importance for the CTS. Therefore, the field of information fusion has to address a broad set of approaches, and should not be reduced to a single topic such as classifier fusion. So far, we have introduced a categorization of fusion algorithms to put of focus on different requirements in a CTS. In the following, we will provide examples for this categorization. Later on, an implementation of larger CTS will be presented and explained. In the literature, such implementations are not very well documented. We hope, that the experience will help interested readers to build new advanced CTS.

The first layer, discussed in Section 2, considers approaches operating on multimodal sensor data which are generally based on late classifier fusion of the same recognized classes, i.e. perception fusion. Section 3 addresses the second layer, i.e. knowledge-based fusion, which focuses on how to incorporate symbolic knowledge to enhance the recognition of more complex classes. Section 4 reports about the third layer, the application-level fusion, which comprises approaches being related to the interaction component on application level. Here, the various information, i.e. implicit and explicit user input, must be correctly combined to realize a successful dialog management and an adaptive user interface. Parts of the presented components are utilized to realize an operating demonstration scenario which is introduced in Section 5. Here a detailed description is provided along with an outline of the experiences gained. Section 6 discusses the architecture, in particular the characteristics and challenges with respect to information fusion. In Section 7, we draw a conclusion and give an outlook to future work.

2. Perception fusion

The perception fusion is performed close to the acquisition of sensor data and has characteristic properties when compared to fusion in higher abstraction levels [28–30]. The input is usually given by decisions of independent base classifiers which work on multiple modalities and recognize the same set of classes. Due to the fact that the system operates in a real-life scenario, it is advantageous that the output of the classifiers consists of probabilistic class memberships. Also the dynamics of the observations have to be captured either with the help of suitable features or by specific classification algorithms. The temporal fusion of the multimodal classifier memberships aims at providing more robust and enhanced predictions which abstract from the sensory for further processing. To allow a consistent handling of uncertainty, the output of the fusion should as well provide a temporal stream of class memberships.

In the following, three approaches for perception fusion are presented: (1) adaptive fusion based on classifiers operating on dynamic features, (2) Kalman filter classifier fusion; and (3) Markov fusion networks.

2.1. Related work

The combination of classifiers has been studied by now for many decades, resulting in a wide spectrum of promising approaches

[48,49,28,50]. Early work on decision combination in multiple classifier systems started in the nineties and represented the classifier output in form of ranks which are fused using rank aggregation methods, e.g. highest rank, Borda count and logistic regression [51,49]. Later on, combination methods such as the sum rule and the product rule became popular. But also alternatives such as the max rule or the median rule were studied in detail [48]. However, so far the work has largely focused on combining samples without temporal extension. In 1999, Jeon and Landgrebe [52] proposed two fusion approaches for multi-temporal classifiers: the *likelihood decision fusion rule* and the *weighted majority decision fusion rule*, which both make use of temporal data to find a global decision. Along with the first approaches in temporal fusion, the field of classifier fusion evolved increasingly fast. New fusion methods, such as Dempster–Shafer combination, fuzzy integrals, decision templates and neural networks come into focus [49,53]. Furthermore, the option to reject samples and to withhold the class assignment due to the lack of confidence in multiple classifier systems has been studied more intensively [49]. At the same time characteristics and design principles are getting analyzed more systematically [54–56]. In 2003, Dietrich [29,30] presented his work on temporal sensor fusion in which he comprehensively investigated early, mid-level and late fusion architectures. Bach et al. [57] proposed a novel kind of weighted feature fusion for support vector machines named multiple kernel learning. The adaptive fusion approach was also pursued by Poh and Kittler who proposed to assess the quality of features and utilize the corresponding measure to weight the classifier system [58]. Recently, the importance of uncertainty in large classifier systems was affirmed by a thorough study carried out by Thiel et al. [50,59].

2.2. Fusion of dynamic features

One of the major challenges in CTS is to provide an interface to the user that allows a natural and anthropomorphic interaction. CTS aim at realizing this feature by recognizing implicit user inputs such as conversational dispositions, behavioral cues and social signals [31,32,4]. These inputs are usually derived by conventional channels such as audio and video, which focus on speech, facial expressions, hand and body gestures, as well as tactile inputs [33–35]. However, especially in case of implicit user inputs, the information to be extracted is superimposed by a large fraction of noise and unrelated signals.

In the example of facial emotion recognition, research in psychology identified that facial expressions are based on action units which disassemble the countenance into separate movements of muscles [36]. According to Ekman et al. [37] these action units give rise to a model that decomposes human sensations into six basic emotions, i.e. happy, sad, disgust, surprise, anger and neutral. Recently developed classifier systems showed that this fundamental research can be successfully put into practice [38,39]. However, the model of six emotions is far to general to provide a viable contribution to the abstract knowledge processing of a CTS. Therefore, alternative models of emotions such as affective states and conversational dispositions gain increasing attention [32,4], albeit they often take place in unrestricted settings. It is worth mentioning that still only few datasets for emotion recognition exist that are designed in such unrestricted settings [40] since recording and ground truth elicitation bear many challenges. For instance, recorded subjects generally tend to show only weak emotions such that a crisp assignment of an emotional state is almost impossible.

In addition to the problem of ground truth elicitation, the unrestrictive setting brings challenges to the training of classifiers and their fusion. For instance, recognizers operating on facial data have to cope with a missing data source (e.g. due to the subject turns away) or problems in feature extraction (e.g. due to the wearing of glasses or a spoken utterance overlaying the facial expression). These kinds of problems often result in a fragmented output of the classifiers over time. Similar observations can be made in the auditory channel. Obviously both channels perfectly complement each other and this yields hope that a multimodal classifier architecture can cancel out parts of the fragmentation.

We consider affective states as parameters of a causal dynamic system consisting of a series connection of user and CTS. In fact, due to the nature of audio and video being functions over time, we obtain time series of features and intermediate classifier decisions. It is obvious that temporal dependencies between different states exist and that the history of previous user states has an influence on the actual state and can thus be exploited by dynamic features [38,33]. However, although features can be designed to capture the information of the target class optimally, they are still overlaid with other signals (e.g. factual content of the utterance) or unrelated noises (e.g. unintended sounds or background noise). In addition, also the target class is characterized by a vast variety of appearances. Our investigations showed that linear classifiers often outperform non-linear classifiers [34,41], since restrictive classifier functions are apparently more robust against noise. Further improvement has been realized by making use of ensembles and exploiting the temporal characteristics. While ensemble approaches help us to capture the variety of the target class [42,43], a time window of features has been used to input the classifiers in order to learn the dynamics of affective states [34].

The presented principles have been examined in the context of a Wizard-of-Oz experiment [44] called the LAST MINUTE corpus [45]. The ground truth of the affective states “normal” (Baseline) and “stressed” (Challenge) is directly given by the screenplay of the

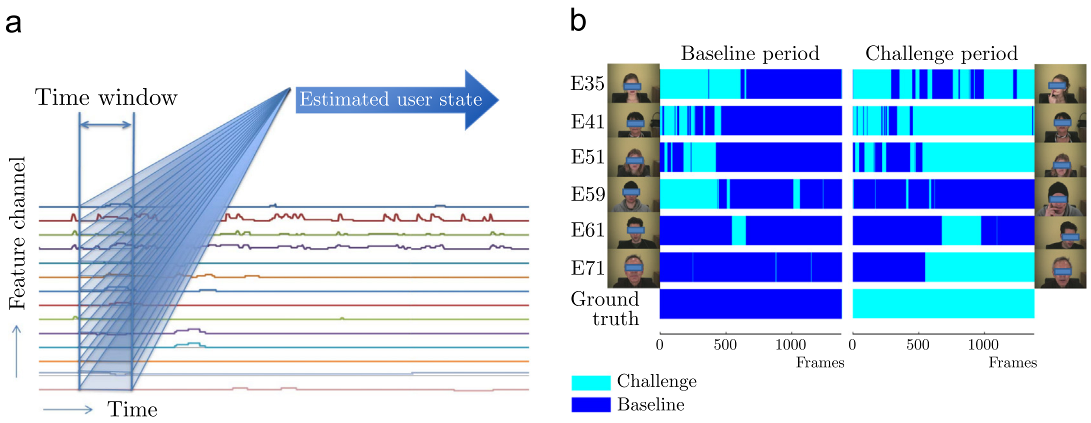


Fig. 3. *Classification of affective states in the LAST MINUTE corpus.* (a) Classification using a shifted window. (b) Classification results. (a) The eye blink frequency and 13 geometric features are collected in a time window to input a linear classifier. and (b) Classification results for selected (anonymized) subjects in Baseline (blue) and Challenge (cyan) period: E35: 61%, E41: 85%, E51: 77%, E59: 33%, E61: 57%, E71: 80%.

experiment such that the need of annotation is avoided. Fig. 3a shows the recognition of the affective state from facial features. The facial features base on the eye blink frequency and normalized geometric distances which are inspired by action units. The linear classifier is operating on a set of features taken from a time window. The weights for the time series are determined in a way typical for matched filter or deconvolution [46,47]. The length of the window, which has been determined using leave-one-subject-out, comprises 15 frames (0.6 s). A detailed evaluation on the influence of the window size is given in [34] and the combination with gestural and prosodic features in [35]. The results over time are shown in Fig. 3b. The unrestricted setting does not ensure that the subject is always in the desired affective state, however, the figure shows clearly that the affective state recognition is capturing the relevant moments in time.

2.3. *Kalman filter for classifier fusion*

Classifiers operating in an unrestricted setting, which additionally have to deal with a low proportion of relevant information, e.g. in the context of emotion, can express their uncertainty by providing an additional confidence measure [60,61]. The probability of a class membership, usually given by $p(\mathrm{y}=y|\mathbf{x}=\mathbf{x})$, where $\mathbf{x}$ is a feature vector and y the target random variable, is already commonly utilized for fusion [62,63]. However, class confidences can provide an additional quantity which is either derived from the individual class membership probability [64] or from the standard deviation of a classifier ensemble [60,43]. These confidence values can be applied in various ways to enhance fusion: on the one hand, confidences can define the influence of a membership in fusion, whereas on the other hand, a low confidence can directly lead to a threshold-based rejection of the classifier decision [65]. In first studies, we make use of the latter concept which is related to sensor failures resulting in missing classifier decisions. However, discarding classifier decisions has the obvious disadvantage that, in case a decision is required by upper layers of the CTS, no decision is available at all. Hence, it is helpful to integrate multiple decisions to recover missing ones and to improve the overall accuracy. Since HCI episodes typically last for a longer time span, the integration of decisions over time is possible and most likely of benefit [60,63,6]. Fig. 4 shows the proposed framework for multimodal classification over time.

A related setting, in which continuous streams of uncertain measurements are used to infer the underlying true quantity, is solved by the well-known Kalman filter [66]. In this setting, we replace the real-world measurements by class membership assignments. The input to the Kalman filter is a temporal sequence of the length T containing M classifier decisions $\mathbf{X} \in [0,1]^{M \times T}$. The inference of a Kalman filter is divided into two steps. In the first step, the belief state is obtained by

$$\hat{\mu}_{t+1} = a \cdot \mu_t + b \cdot u \tag{1}$$

$$\hat{\sigma}_{t+1} = a \cdot \sigma_t \cdot a + q_m \tag{2}$$

where all quantities of the regular Kalman filters are replaced by scalars. In Eq. (1), the state transition model a and the control-input model b weight the previously obtained mean μ_t and the control u linearly. Eq. (2) represents the uncertainty with respect to state $\hat{\mu}_{t+1}$. The control u can be used to bias the prediction to a certain value (e.g. the least informative classifier combination outcome 0.5 in case of a two-class problem ranging between $[0,1]$). However, we decided to omit the last term of Eq. (1) such that our model presumes that the mean of the current estimate is identical to the previous one. Due to the restriction of the state space to the values $[0,1]$ in classifier fusion, the usage of popular process models like dead reckoning, which propagates the state using the last state and its first derivation with respect to the time, is not possible. Alternatively, a non-linear version of the dead reckoning model would be necessary to keep the state restrictions. The covariance of the prediction is given by $\hat{\sigma}_t$ and obtained by combining the *a posteriori* covariance with an additional covariance q_m which is the process noise. The successive update step is performed for every classifier m with the corresponding decision $y_{m,t+1}$ and requires three intermediate results, namely the residuum γ, the innovation variance s and the Kalman gain k_{t+1}:

$$\gamma = y_{m,t+1} - h \cdot \hat{\mu}_{t+1} \tag{3}$$

$$s = h \cdot \hat{\sigma}_{t+1} \cdot h + r_m \tag{4}$$

$$k_{t+1} = h \cdot \hat{\sigma}_{t+1} \cdot s^{-1} \tag{5}$$

where h is the observation model, which maps the predicted quantity to the new estimate and r_m is the observation noise. These outcomes are then used to obtain an updated mean and variance:

$$\mu_{t+1} = \hat{\mu}_{t+1} + k_{t+1} \cdot \gamma \tag{6}$$

$$\sigma_{t+1} = \hat{\sigma}_t - k \cdot s \cdot k. \tag{7}$$

Missing classifier outcomes (decisions) are replaced by a measurement prior $\tilde{y}_{mt} = 0.5$ and a corresponding observation noise $\tilde{r}_m$ which is set relatively high compared to the actual observation noise. Fig. 5 shows the output of the Kalman filter, combining the decisions of two modalities for the class "arousal" [41], from a study conducted using the AVEC 2011 dataset [67,68]. The recognition results of the audio (blue squares) and video (orange dots) channel are plotted along the time axis (only one probability mass of the binary classification result is shown). Results with low confidences are rejected, and therefore plotted in a pale color. The solid black line represents the estimate of

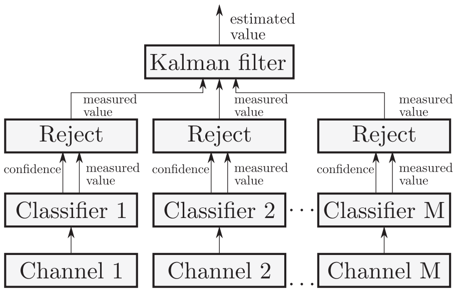


Fig. 4. *Multiple classifier system utilizing a Kalman filter to combine classifier decisions.* Based on independent features, a set of M classifiers produce a temporal stream of predictions and confidences. The confidences are utilized to reject weak classifier decisions. Missing decisions due to rejection or sensor failures can be recovered by the Kalman filter.

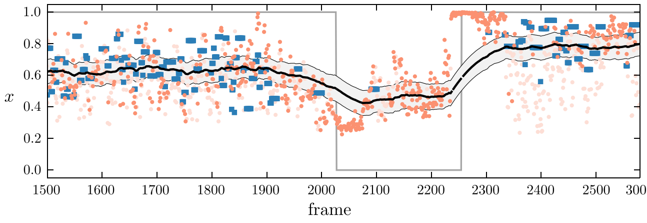


Fig. 5. *Kalman filter based fusion of multiple modalities using data from the AVEC 2011 dataset targeting the class "arousal"*. Orange dots and Blue squared-shaped markers correspond to the video and audio decisions, respectively. Markers in pale color do not contribute to the fusion. The thick black curve corresponds to the fusion result, while the area around the curve corresponds to the variance of the Kalman filter (scaled by 10 for illustration purposes). The light gray curve displays the ground-truth. The parameters used are $q_{\mathrm{audio}} = 10^{-6}$, $q_{\mathrm{video}} = 10^{-5}$ and $r_{\mathrm{audio}} = r_{\mathrm{video}} = 0.75$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 1
Accuracies and F$_1$-measures on the AVEC 2011 dataset utilizing Kalman filter classifier fusion and the reject option and unimodal results. The F$_1$-measure is defined by $F_1 = 2P \cdot R/(P+R)$ where P is the precision and R the recall.

(a)

Online	Arousal	Expectancy	Power	Valence
↑ Accuracy	68.5(5.7)	62.5(4.9)	61.8(6.6)	64.2(9.3)
↑ F_1	72.6(4.2)	42.2(15.7)	69.1(7.6)	72.6(10.9)
↑ $\overline{F}_1$	59.7(15.1)	71.1(5.8)	43.5(18.0)	43.7(3.0)
Audio reject	0%	0%	0%	90%
Video reject	50%	50%	0%	10%

(b)

Audio	Arousal	Expectancy	Power	Valence
↑ 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	**Arousal**	**Expectancy**	**Power**	**Valence**
↑ 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)

the Kalman filter, whereas the gray line depicts the ground truth [41]. An excerpt of the statistical evaluations, performed on a re-partitioned dataset containing subject independent folds, is shown in Table 1. The best performing submissions to the original challenge achieved an accuracy of slightly above 60% for the class “arousal”, the three classes, i.e. “expectancy”, “power” and “valence”, have been withdrawn from being rated. Table 1a shows the audio-visual classification performance of the Kalman filter classifier fusion, whereas Table 1b shows the corresponding uni-modal results without rejection.

2.4. Markov fusion networks for classifier fusion

Late fusion algorithms being deployed in real-time cognitive technical systems have to be suited to a wide range of problem specific conditions. First of all, the algorithms have to perform close to real-time and be capable to combine decisions from multiple sources and, most importantly for CTS, of multiple classes. Furthermore, the algorithm needs to be tolerant against sensor failures while being able to deal with a high degree of uncertainty arising from the open world and possibly weak labels (depending on the classification problem addressed). However, although there are many problems, the given setting provides also opportunities for enhancement, e.g. taking temporal information or symbolic knowledge into account. Within this section a novel fusion technique is presented, which is well suited to meet these requirements, namely the Markov fusion network (MFN) for late temporal classifier fusion [6]. Likewise to the Kalman filter fusion approach each modality m provides a stream of class distributions in form of a matrix $\mathbf{x}_m \in [0,1]^{I \times T}$ where T denotes the length of the stream and I the number of classes to be recognized. The output of the algorithms is a fused estimate $\mathbf{Y} \in [0,1]^{I \times T}$.

The MFN is defined by three potentials being part of the energy function to be optimized, namely the *data potential* Ψ, the *smoothness potential* Φ and the *distribution potential* Ξ. The data potential Ψ enforces the estimate $\mathbf{y}_t \in [0,1]^I$ to be similar to the input class distribution $\mathbf{x}_{mt} \in [0,1]^I$ of the mth source and is defined by

$$\Psi = \sum_{m}^{M} \Psi_m$$

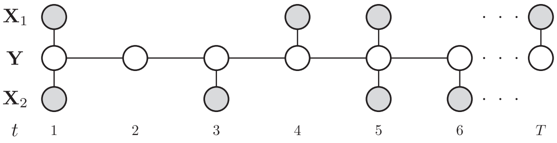


Fig. 6. *Graphical model of the Markov fusion network (MFN).* The MFN combines classifier decisions of multiple modalities and with respect to its temporal relationship. The estimated decisions $\mathbf{Y}$, which may be composed of multiple classes, are temporally connected using a Markov chain and additionally influenced from available decisions of the individual modalities, here $\mathbf{X}_1$ and $\mathbf{X}_2$.

$$= \sum_{m}^{M} \sum_{i=1}^{I} \sum_{t \in \mathcal{L}_m} k_{mt}(x_{mit} - y_{it})^2 \tag{8}$$

where $\mathbf{K} \in \mathbb{R}_+^{M \times T}$ individually rates the reliability of the classifier m at time step t. The set $\mathcal{L}_m$ contains the time steps not affected by sensor failures. The second potential Φ models the Markov chain enforcing lateral smoothness and is given by

$$\Phi = \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{\hat{t} \in N(t)} w_{\min(t,\hat{t})}(y_{it} - y_{i\hat{t}})^2 \tag{9}$$

where $\mathbf{w} \in \mathbb{R}_+^{T-1}$ weights the difference between two adjacent nodes in the chain and $N(t)$ returns the set of adjacent nodes, e.g. $N(t) := \{t-1, t+1\}$ in case both neighbors are available. The third potential Ξ asserts that the resulting estimate is conformed to the laws of probability theory and is given by

$$\Xi = u \cdot \sum_{t=1}^{T} \left(1 - \sum_{i=1}^{I} y_{it}\right)^2 + \sum_{i=1}^{I} y_{it}^2 1_{[0 > y_{it}]} \tag{10}$$

where u weights the relevance of the potential and $1_{[0>y_{it}]}$ is one in case y_{it} is negative. The potential enforces the estimate to sum up to one for each time step and penalizes negative values. Fig. 6 shows a sample graphical model fusing two modalities ($M=2$). The estimate $\mathbf{Y}$ to be derived is represented by the Markov chain of white nodes. For each time step, the gray nodes of the classifier input distribution $\mathbf{X}_1$ and $\mathbf{X}_2$ are connected to the corresponding estimates. In case no distribution is available, e.g. due to sensor failure or the rejection of results with low confidences, the nodes are omitted [6,69,35]. Real-time fusion can be realized by shifting a window over the input data. Fig. 7 shows the output of the MFN using the same data as the fusion using the Kalman filter, being

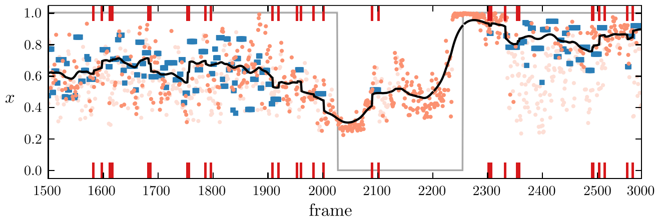


Fig. 7. *Markov fusion network combination of multiple modalities using data from the AVEC 2011 dataset targeting the class* "*arousal*". The orange dots and the blue squared-shaped markers correspond to the video and audio decisions, respectively. Markers in pale color have been rejected and do not contribute to the fusion. The red ticks indicate turns or pause within the conversation. The model is based on the assumption that pauses and turns give evidence for a change in emotion (or the other way round: remains stable within one conversational episode). The thick black curve corresponds to the fusion result. The light gray curve displays the ground-truth. The parameters used are $w_{normal} = 128, w_{turn} = 4, k_{\text{video}} = .5$ and $k_{\text{audio}} = 5$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 2
Accuracies and F_1-measures on the AVEC 2011 dataset utilizing Markov fusion networks and the reject option of confident classifier decision.

Online	Arousal	Expectancy	Power	Valence
↑ Accuracy	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)
↑$\bar{F}_1$	63.4(8.6)	69.6(5.2)	45.7(14.9)	40.3(10.3)
Audio reject	10%	50%	90%	90%
Video reject	50%	90%	0%	90%

depicted in Fig. 5. The red ticks on the top and bottom of the figure correspond to conversational turns in which the temporal smoothness is suppressed based on the assumption that an affective state change remains stable within a turn. The modification leads to the formation of stable regions and spots of changes whenever a turn occurs.

Table 2 shows the statistical evaluation in the same settings as described in Section 2.3 [70]. The accuracies of the classes "arousal" and "valence" perform slightly better. However, the changes are only marginal which demonstrates that both techniques have a great potential. Due to the individual weighting of the modalities, the shares of rejected decisions show significant differences. However, since these quantities have many dependencies only limited information about the characteristics of the input modalities can be inferred.

3. Knowledge-based fusion

In knowledge-based fusion the classes recognized by the perception-level fusion are enriched with models created by human experts. Generally, the aim is to obtain a more abstract or contradiction-free representation of the input [71,72]. The high variability of complex patterns makes the recording of sufficiently large datasets for classical machine learning algorithms infeasible. However, these complex patterns can often easily be decomposed into smaller sub-patterns which in turn are very well suited for machine learning algorithms. Hence, it is obligatory that at some level of abstractness the recognition has to switch over from robustly trained patterns to approaches from symbolic artificial intelligence (AI). However, the change towards abstract classes has to be performed smoothly since the recognition of sub-patterns is generally afflicted with a high degree of uncertainty. Probabilistic symbolic AI approaches such as Markov logic networks (MLN) [73] or partially observable Markov decision processes (POMDP) [74] for problem sets of a manageable size are therefore advisable. To dissolve contradicting class assignments it may even be beneficial to use alternative probability theories such as the Dempster–Shafer theory of evidence (DST) [75,76]. The DST is applied at multiple location in this paper such that it is reasonable to give a brief introduction to the theory. The theory extends the classical probability by assigning degrees of beliefs to elements of a power set 2^{Ω} where the frame of discernment (FOD) Ω contains atomic hypotheses. Degrees of beliefs are modeled by the mass function $m: 2^{\Omega} \rightarrow [0,1]$ where $\sum_{\mathcal{A} \in 2^{\Omega}} m(\mathcal{A}) = 1$. An assignment to $m(\varnothing)$ describes the special case in which none of the hypotheses are supported (open world assumption). On the other side, the assignment to $m(\Omega)$ represents the ignorance and that less credibility is put to a specific hypothesis. The combination of two mass functions $m = m_1 \bigcirc\!\!\!\!\!\cap\, m_2$ can be performed using Dempster's (unnormalized) rule of combination:

$$m(\mathcal{C}) = \sum_{\mathcal{C}:\mathcal{C} = \mathcal{A} \cap \mathcal{B}} m_1(\mathcal{A}) \cdot m_2(\mathcal{B}) \tag{11}$$

in which the mass function given the empty set can take values bigger than zero. In case the open world scenario is not an option, the combination of two mass functions $m = m_1 \oplus m_2$ can be performed using the normalized Dempster's rule of combination:

$$m(\mathcal{C}) = \begin{cases} \dfrac{(m_1 \bigcirc\!\!\!\!\!\cap\, m_2)(\mathcal{C})}{1-K} & \text{if } \mathcal{C} \neq \varnothing \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where $K = (m_1 \bigcirc\!\!\!\!\!\cap\, m_2)(\varnothing)$ describes the degree of conflict [75,76].

Three approaches are presented to exemplify the knowledge-based fusion: (1) knowledge-based sensor fusion in tracking, (2) track-person association using MLN; and (3) complex class recognition based on layered Markov models.

3.1. *Related work*

Knowledge-based information fusion aims at integrating sub-symbolic and symbolic approaches in order to improve the overall classifier system's performance. Before going into details, it is mandatory to provide an intuition of what is meant by the terms symbolic and sub-symbolic, since it is tempting to relate these terms to the information which serve as input. However, it is possible to interpret data as either being sub-symbolic or symbolic, e.g. color information of an image. Despite the appearance of the data, we therefore emphasize that the manner of processing is decisive for the categorization of the symbolic and sub-symbolic approaches [71]. The early symbolic approaches of AI which aimed at modeling the mental function of human showed clear limitations (e.g. the missing ability to store and access a huge quantity and variety of knowledge, the lack of an elaborated ability for recognition in many domains and the difficulty to find new relations between pieces of information [77]). To overcome these limitations, researchers started to combine sub-symbolic and symbolic approaches. First approaches aimed at transferring the techniques from AI directly to the connectionist's setting [77,78]. However, in general the intended improvement in performance could not be achieved and these approaches have only few offsprings.

In contrast to the ambitious project of modeling human mental functions, other approaches pursuit a more basic goal, namely the combination of information sources in a specific domain in order to recognize more abstract classes or to dissolve contradicting source-specific recognitions. So far, not much work has been done in this field. Wrede et al. [79] developed a cognitive vision system for scene analysis which makes use of an active memory. The system is capable of learning objects and actions of a scene and stores the concepts in a volatile memory. The content of the memory is processed by additional algorithms, e.g. a consistency validation, which would reject the recognition of certain actions in case no supporting objects are found in the scenery. An approach to recognize office activities, which takes the context of manipulated objects into account, has been proposed by Biswas et al. [80]. The information of the context is modeled using an MLN [73] which operates on multiple weak feature sources, i.e. object information, body pose and body movement. The combination of the observed features using the MLN resolves the ambiguity and results in a reliable recognition. Tran and Davis [81] presented a system which enhanced the surveillance of an outdoor parking lot utilizing a MLN [73]. Despite the noisy observations, the system recognizes humans and cars to derive high-level information with the help of a MLN, e.g. people shaking hands, trunk get loaded or a person enters a car. Tenorth and Beetz [82] proposed a knowledge processing system based on first-order logic to enable robots to act autonomously. The system can not only recognize objects stored in the ontology but can also infer the possible locations of objects based on their functionality and can automatically complete underspecified instructions of humans. Kembhavi et al. [83] developed a related system for scene understanding, which in contrast utilizes a probabilistic MLN and has a closer integration of image analysis and reasoning. The goal is to assign a functionality to scene elements which have been generated during a preprocessing segmentation step, e.g. “road”, “sidewalk”, “crosswalk”, “pedestrian entrance” or “bus-stop”. The input to the MLN is given by the segmented zones, the tracked objects over time, e.g. “bus”, “car” or “pedestrian”, and the common world knowledge in the form of rules, such as “cars stop at crosswalk in order to let people pass”. Gehrig et al. [84] utilize a multi-level approach [85] to detect nine human intentions (e.g. “prepare cereals”, “prepare pudding”, “clear table”) based on six activities (e.g. “prepare meal”, “clear table”) which are again composed of a set of 60 motion primitives (e.g. “place object on table”, “pour” or “stir”). Domain knowledge is applied (by the means of ground truth data to learn transition probabilities) either to the motion recognition, to the activity recognition, or both, in order to improve performance. In summary, it can be said that although not much work can be found in the literature yet, a growing interest has developed over the recent years [86,87].

3.2. *Multi-object tracking incorporating inter-object state dependencies and sensor specific properties*

CTS need to have a dynamic model of the environment to adapt to the latest situation. This dynamic model can be realized following the concept of multi-object multi-sensor fusion algorithms which can provide states of all objects in the environment. Representatives of this concept are multi-object Bayes (MOB) filters [88] which are based on random finite set statistics. By using random finite sets, a state of the MOB filter represents the complete environment. Thus, in contrast to standard tracking algorithms using multiple Kalman filter instances, the MOB filter allows for the incorporation of inter-object dependencies.

In [89,90], we applied the MOB filter to the scenario of CTS and presented the first real-time capable implementation of the filter. In addition to the assumed motion model of the persons, the prediction step incorporates the usage of a state dependent detection probability and interactions between the objects [91]. The state dependent detection probability of an object is calculated using the sensor properties and the state estimate of all other objects within a realization of a random finite set. The interactions are used to integrate a preferred distance to other persons in the prediction step and to avoid physically impossible multi-object states.

The performance gain obtained by modeling interactions between objects is illustrated using a split and merge scenario. First, the two objects are well separated in this scenario. Until $k=60$ the objects are approaching until they are next to each other. Finally, at $k=100$, the objects start separating again. The performance of the sequential Monte Carlo implementation of the MOB filter (SMC-MOB) is compared to the joint integrated probabilistic data association (JIPDA) filter [92] using the optimal subpattern assignment metric for tracks (OSPAT) [93]. The OSPAT represents the deviation between true and estimated state of an object and the difference between estimated and true number of objects by a single scalar value. Additionally, a parameter α is used to penalize different track IDs for a single object. For $\alpha=0$, ID switches are discarded and the OSPAT distance can be interpreted as the average distance per object between the true and the estimated positions. If α equal the cut-off value c, and ID switch is penalized like a missed detection of the object.

Fig. 8 shows the OSPAT for the split and merge scenario. The OSPAT distance of the JIPDA filter ascends immediately after $k=60$, i.e. when the objects are very close to each other. On the other hand, the SMC-MOB filter only has minor difficulties to handle the closely spaced objects between $k=60$ and $k=100$ which results in a nearly constant OSPAT value during this period of time. When the objects start separating again, the JIPDA filter has a much higher possibility of track ID switches as the SMC-MOB filter, which result in a larger increase of the OSPAT distance. Thus, the ability of the SMC-MOB filter to model object interactions results in a measurable performance gain compared to the JIPDA filter. An application of the interaction model to conventional tracking algorithms like JIPDA is not possible, since each object is tracked using an object-individual Kalman filter instance.

In multi-sensor fusion, the different perception properties of the sensors are challenging. On the one hand, video sensors show excellent performance in detecting objects of specific types, e.g. pedestrians, using cascaded classifiers [94] but they are not able to distinguish between other object types and empty space. On the other hand, laser range finders can easily separate between empty space and occupied areas, but object classification has high false positive probabilities since different object types may have similar shapes. Consequently, the different perception properties lead to contradictory measurements if an object type is not detectable by the video sensor.

In [95], we propose to handle this behavior using the DST. The DST is used to model the perception properties of the sensors, i.e. a laser range finder and a video camera, to resolve the contradiction. The frame of discernment (FOD) Ω consists of the atomic hypotheses relevant object (R), other object (O), and no object (N). Assuming that a detection in the camera image has a true positive probability of p_{TP}, its mass distribution is given by

$$m_C(R)=p_{TP}, \tag{13}$$

$$m_C(NO)=1-p_{TP}. \tag{14}$$

As mentioned above, a laser range finder can rarely distinguish between O and R. Thus, its mass distribution is given by

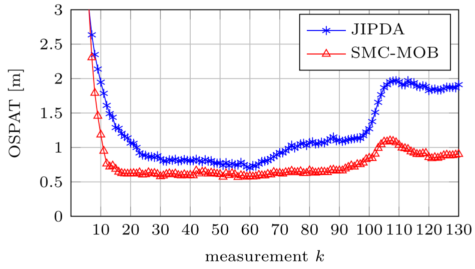


Fig. 8. *Comparison of the SMC-MOB and the JIPDA filter.* Mean value of the OSPAT distance over 100 Monte Carlo runs. The order of the metric is $p=1$ and the cut-off value is $c=10$. Using the parameter $\alpha=c$, switching track IDs are penalized like missed detections.

$$m_L(OR) = p_{TP}, \quad (15)$$

$$m_L(N) = 1 - p_{TP}, \quad (16)$$

which correctly models the classification uncertainty. Fig. 9 shows an example for an update using the frame of perception approach. Combining the measurements without modeling the frames of perception using DST, the incorporation of the missed detection of the video would result in a higher evidence for the hypothesis that the object does not exist. The JIPDA tracker with DST based measurement modeling outperforms the algorithm without DST by means of a lower false alarm rate for equivalent detection rates [96], which corresponds to an increase of the area under the ROC curve by 9.9%.

The incorporation of inter-object dependencies in the MOB filter as well as the DST based measurement modeling significantly improves the tracking results. Thus, a combination of the DST based measurement modeling with a MOB filter is supposed to provide further improvements of the tracking results and will be investigated in the future.

3.3. Track-person association using a first-order probabilistic model

On a higher level of information processing it is advantageous to include symbolic knowledge, which may either be generated directly, for example from dialog interactions with a user, or is the result of cognitive reasoning processes of the CTS itself. A prominent source for high-level knowledge that is generated within the system is predictions on future trajectories as the result of automatic planning [97,98]. Since the planning processes need to integrate information about the dynamic behavior of the environment and high-level information about the user's behavior and his goals, the value of information gained from planning can be of great importance.

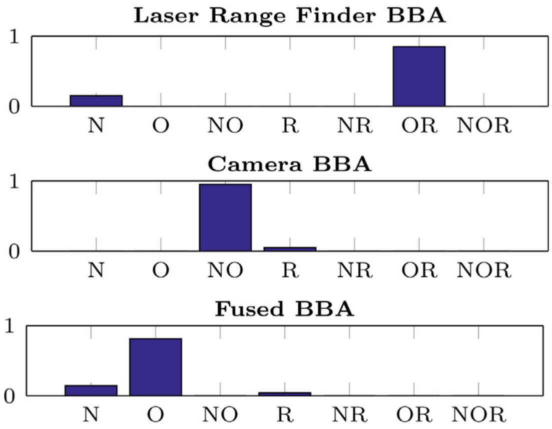


Fig. 9. *Update using frames of perception.* The laser measurement (top plot) is not able to distinguish between O and R, whereas the camera BBA (middle plot) indicates a missed detection in the video. The BBA after combining the two sensors indicates that the object is of type O (where N represents the non-existence of the object).

However, the character of such knowledge is often very different from sensory data. Automated planning solves more abstract problems that are often specified by human experts in a first-order language. These models often contain deterministic dependencies, as they are usually formulated in a logic-like language [99,100]. The task will be to bridge the gap between these first-order symbolic models and the perception models that usually deal with probabilistic and continuous data.

First-order probabilistic languages [101–103] try to offer the best of both worlds, since they allow the construction of relational probabilistic models. We chose Markov logic networks (MLN) [73] as a tool to integrate tracking information (obtained by the multi-object tracking approach described in Section 3.2), static environmental knowledge and information about users' goals [104,9].

The multi-object tracking approach can yield excellent results on its own in situations with low occlusion. However, as soon as several persons are within the scene at the same time, an object tracker based on data from a laser range finder alone begins to confuse the track associations. In order to recover the correct association of tracks to persons within the scene we constructed a MLN model that successfully integrated partial information about persons' destinations with a map of the environment and the probabilistic tracking information [104].

Since MLN base their semantics on discrete Markov networks, a practical discretization scheme had to be developed to process the continuous output generated by the MOB. Maps of different granularity were evaluated for binning the particles generated by the filter of the MOB into discretized regions [9]. One such discretization scheme is depicted in Fig. 10.

The MLN consists of logic-like rules that make probabilistic propositions. For example the following rule taken from the MLN describes the fact that if a track m is at a certain position r and this track is associated to a person p, then that person must be at the same position at the same time t:

$$\texttt{TAt}(t,m,r) \wedge \texttt{assoc}(m,p) \Rightarrow \texttt{PAt}(t,p,r). \quad (17)$$

The formula creates a dependency between the three predicates: (1) TAt, which represents the position of tracks; (2) PAt, which represents the position of persons; and (3) assoc, which represents the association between a person and a track. The formula in the used model is of probabilistic nature and allows for some slip to occur

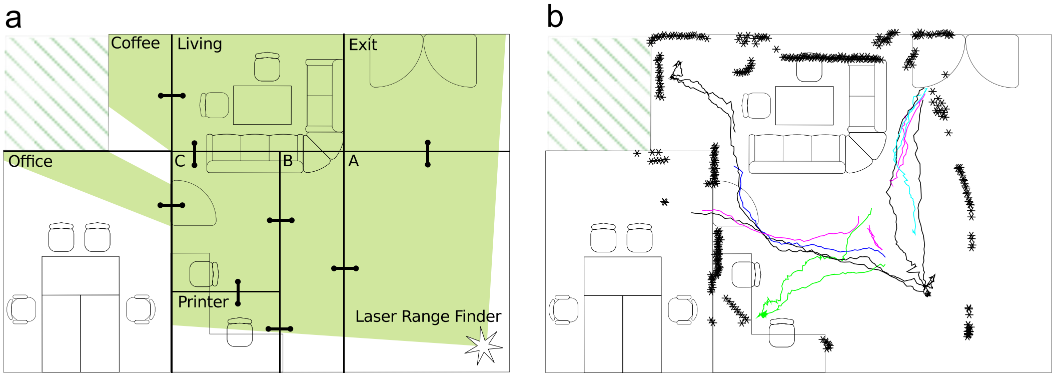


Fig. 10. *Multi-object tracking integrating first-order symbolic models and the sensory data.* (a) Possible discretization scheme for the continuous tracking data. The handles represent adjacency of regions. The adjacency is used in the MLN model to limit the movement of the objects within the scene. The green area is covered by the laser range finder. (b) The lines signify the most probable trajectories of two persons moving simultaneously within the scene, as detected by the multi-object tracker. Different colors represent five different tracks spawned by these persons. Black crosses represent background measurements of permanent obstacles. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

between the track and its true position. While MLN do not have an explicit notion of time, the two predicates describing locations are turned into dynamic features by making them depended on time. This makes their truth value changing over time, while the association between tracks and persons is stable. The approach has been evaluated using nine sequences in the scenario of Fig. 10 (three sequences with one person, three sequences with two persons, and three sequences with three persons). Each sequence has a duration of about one minute. The standard tracking model, which uses only the basic floor plan, ends up with 13 false associations over all tracks. A second model which uses an additional static occlusion model reduces the number of false association down to 9. The final model using, which further brings in information about personal goals as proposed in this section, has only 4 false associations. A detailed description of the study can be found in [104].

In general, one can say that using first-order probabilistic languages enables creating models which are partly or as a whole specified by hand, which make the usage of available background knowledge easy and reduce the amount of data required to train model's parameters. But the created models are often of a high complexity and inference can become quickly computationally infeasible.

3.4. Layered Markov models for complex class recognition

The recognition of complex patterns is crucial for CTS to provide meaningful events to the system's inner building blocks, as for instance the knowledge base or dialog management. However, the recognition of patterns, like “having a breakfast” or “working”, bears many challenges, since it requires to make use of temporal and symbolic knowledge. On the other hand, complex patterns can often be decomposed into sub-patterns which are generally easier to be recognized. In case of emotion recognition, basic observable behavioral cues, e.g. to turn away or to cut into an ongoing dialog, can be used to derive more complex patterns, such as the current affective state or conversational dispositions [4]. Furthermore, the concept of temporally composed patterns can be pursued to enable the systems to generate a long-term user categorization.

In 2004, Oliver et al. [85] proposed the utilization of layered hidden Markov models (HMM) to compose complex patterns out of smaller sub-patterns [8]. We extended the concept of layered HMM to a generic layered model, which not only generalizes the classifier models but also incorporates independently derived symbolic knowledge, e.g. by making use of conditioned HMM (CHMM) [105–107]. Fig. 11 depicts the functional principle of the architecture. In the first layer, the raw input data is passed to a sequential classifier, e.g. Markov models in which the classification of basic patterns is performed. For each time step, the outputs are collected such that a new stream of data is generated. The second layer then uses the new stream as input to the next classifier by shifting another window. This number of layers varies depending on the problem setting. The time granularity and the abstractness of the classes increase in the upper layers, because the classes to be detected are composed of the underlying layer's classes. As a result, the recording of training material for the upper layers is associated with a bigger effort, for the duration and the variability of class occurrences increase. This issue is addressed by introducing the MLN [73] to the architecture [8]. With the help of MLN, high-level knowledge can be implemented into the recognition architecture via probabilistic logical rules.

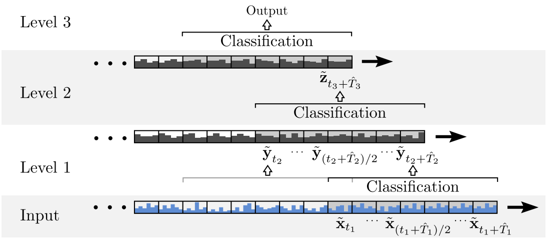

Level 3
Output
Classification
Level 2
$\tilde{\mathbf{z}}_{t_3+\hat{T}_3}$
Classification
Level 1
$\tilde{\mathbf{y}}_{t_2}$ ⋯ $\tilde{\mathbf{y}}_{(t_2+\hat{T}_2)/2}$ ⋯ $\tilde{\mathbf{y}}_{t_2+\hat{T}_2}$
Classification
Input
$\tilde{\mathbf{x}}_{t_1}$ ⋯ $\tilde{\mathbf{x}}_{(t_1+\hat{T}_1)/2}$ ⋯ $\tilde{\mathbf{x}}_{t_1+\hat{T}_1}$

Fig. 11. *Functional principle of the generic layered architecture.* Within each layer, a window is shifted to extract the input for sequential classifiers. The result of the classifier of each shift is collected to form a new stream which is then provided to the next subsequent layer. Each layer is trained on top of the preceding layer such that the time granularity of the pattern and the abstractness of the recognized patterns increase.

The architecture has been evaluated in the area of activity recognition and user categorization. Activities, e.g. “drinking from a cup” or “writing a note”, are decomposed into actions such as “pick up object”, “manipulate object” or “move object towards head”. Additional knowledge about the object class is then used to successfully recognize the activity. On top of the activity recognition, the user categorization is obtained from an MLN which derives the type of the user. The input to the MLN may be any sequence of the known set of activities. In order to keep the first experiment feasible, we categorized the user prefers “black”, “white”, “sweet” or “sweet white” coffee [8,106]. Three different MLN have been evaluated to analyze the third layer. The first model contains the basic rules and introduces a smoothing over time. The second model abstracts from the temporal extends and introduce predicates which indicates whether activities have been already performed or not. The third model extends the second model by incorporating additional domain knowledge such as the user is supposed to drink only from a stirred cup in case milk or sugar has been added. While the first model has an average error rate which is below chance (four classes), the error rate of the second model achieves 41.9%. The additional domain knowledge of the third model further reduces the error rate down to 29.4% [108]. However, recent result on the same dataset using additional data shows even far more promising results. A new experiment, which is currently being analyzed, builds on the insights gained on propagating information upwards to detect more complex classes but aims at propagating information downwards as well in order to improve the recognition results of the lower layers.

4. Application-level fusion

The application-level fusion combines abstract information from multiple sources, namely the explicit user input, the knowledge base (which provides the current and pasts implicit states of the user), the dialog management and the actual application. However, in contrast to the preceding two fusion categories, which have a rather passive way of perception, the inferred information of the application-level fusion is usually directly involved in a decision making process. Since these decisions strongly affect the experience the user gets from the system, incorrect decisions have to be avoided at all costs. Hence, the crucial question is how to handle uncertainty and contradicting facts at this abstract level. Furthermore, the interplay of the components must allow an almost unrestricted dynamic interaction with the system without losing the user's goals.

Within this section three topics are addressed: (1) how does the application and the knowledge base can help to correctly interpret the user input and provide an abstract representation of it; (2) how does fission generate an adequate output based on fused abstract information; and (3) how can the dialog management choose a strategy to react in the probably most appropriate way given a large number of constraints, goals and recently acquired information.

4.1. Related work

The development of user interfaces aims at providing more natural and effective ways of HCI. In early days, the keyboard was the only input to the system, which was then revolutionized by the computer mouse. Nowadays, a multitude of communication channels to a computer system are available, e.g. touch, gestures or speech. However, the user still has to learn how to utilize these inputs rather than communicating naturally and in a unconstrained manner. Up to now, HCI has mostly focused on explicit interaction. Possible implicit input cues remain commonly unnoticed. Furthermore, current systems usually do not combine these multiple modalities, although this feature can be considered as crucial for a CTS since it allows the user to have a more natural interaction [109].

The idea of using gestures and voice for control had first breakthroughs in the seventies [110–112]. In [113], Wahlster proposed a multimodal interface in which user and discourse models are additionally integrated to allow speech references to regions of a form by tactile gestures. Bangalore and Johnston [114] aimed at creating a combined semantic representation for the multimodal utterance with the help of a finite-state automaton such that the gestural information directly influence the recognizer's search. Kaiser et al. [115] proposed an approach to integrate speech and gesture (i.e. point, push twist and rotate) input together with the looking direction for virtual reality object manipulation. Commands follow a grammar and all measurements of the system, e.g. speech, gestures and objects looked at, are associated with a degree of uncertainty. The most probable command is then determined by evaluating the multimodal command with highest score according to the mutual disambiguation rank [116].

Dialog management systems provide abstract interfaces to the user which are adapting to the current situation by having a model of the user, the environment and the application. Most works in this field aim at extracting the correct and most crucial information to advance the discourse efficiently as expected by the user. Generally, information is updated after a conversational turn has been taken [117]. A multitude of dynamical models have been studied so far, ranging from plan-based agents [118] to probabilistic approaches, such as POMDP [119,120]. Young et al. [121] proposed a POMDP which models the information state by partitioning the domain hierarchically in a tree-like structure. Each time when new information is available, the knowledge tree is refined or expanded by distributing belief masses to the nodes. Nguyen and Wobcke [122] proposed a multimodal dialog manager which processes information using four groups: (1) conversational act determination and domain classification; (2) intention identification; (3) task processing; and (4) response generation. Each group is represented by a plan which helps us to derive or generate the intended functionality and thus, follows a unified design pattern.

According to [123], systems combining different output modalities, such as text and speech, have been evolved since the early 1990s. The allocation of output modalities of these early multimodal systems was rather hard-coded than based on intelligent algorithms. An important survey on multimodal interfaces, principles, models, and frameworks is provided by [126]. Beyond that, [126] mentions the idea of machine learning approaches for multimodal interaction. The given example focuses on machine learning in multimodal fusion on the feature level; but such techniques may also be appropriate for fission approaches. Another interesting approach is presented in [127]. The authors present a multi-agent system, in which past interactions are taken into account to reason about the new output, using a machine learning approach for case-based reasoning. To reason about the best UI configuration in a certain context of use (CoU) is a challenging task. Some approaches provide a meta UI where the user can specify a certain UI configuration, e.g. via an additional touch device [128]. Based on that, the system is able to respect the user's demands and can distribute the UI via the referenced device components.

4.2. Multimodal input fusion using the transferable belief model

The task of the multimodal input fusion is to derive a consistent abstract meaning of the observed explicit user inputs. The fusion uses the transferable belief model (TBM) [129] based on DST of evidential reasoning as formal principle, already introduced in Section 2. The rationale for this is that testimonies from “experts” (the lower level fusion components) have to be judged. Therefore, instead of classical probabilities, beliefs are assigned to the events arriving from the lower levels. Such beliefs allow explicit representation of ambiguities in the form of disjunctions like “event A or event B happened with a belief of m”, without the necessity to assign probabilities to individual events.

Using this theory, evidences from different sources about events, described as sets in the FOD Ω that is part of the knowledge base of a CTS, can be combined. In order to actually fuse events (not only beliefs), the FOD is extended by the notion of tuples that denote the occurrence of a combined input. For example, if the user wants to select an object o using a verbal deictic reference like “this one” and a pointing gesture, this would lead to beliefs about a=‘select’ and b=‘object o’. If the FOD contains the tuple (a,b), a belief about the occurrence of the combined event ab=‘select object’ can be computed using the following modified rule of combination:

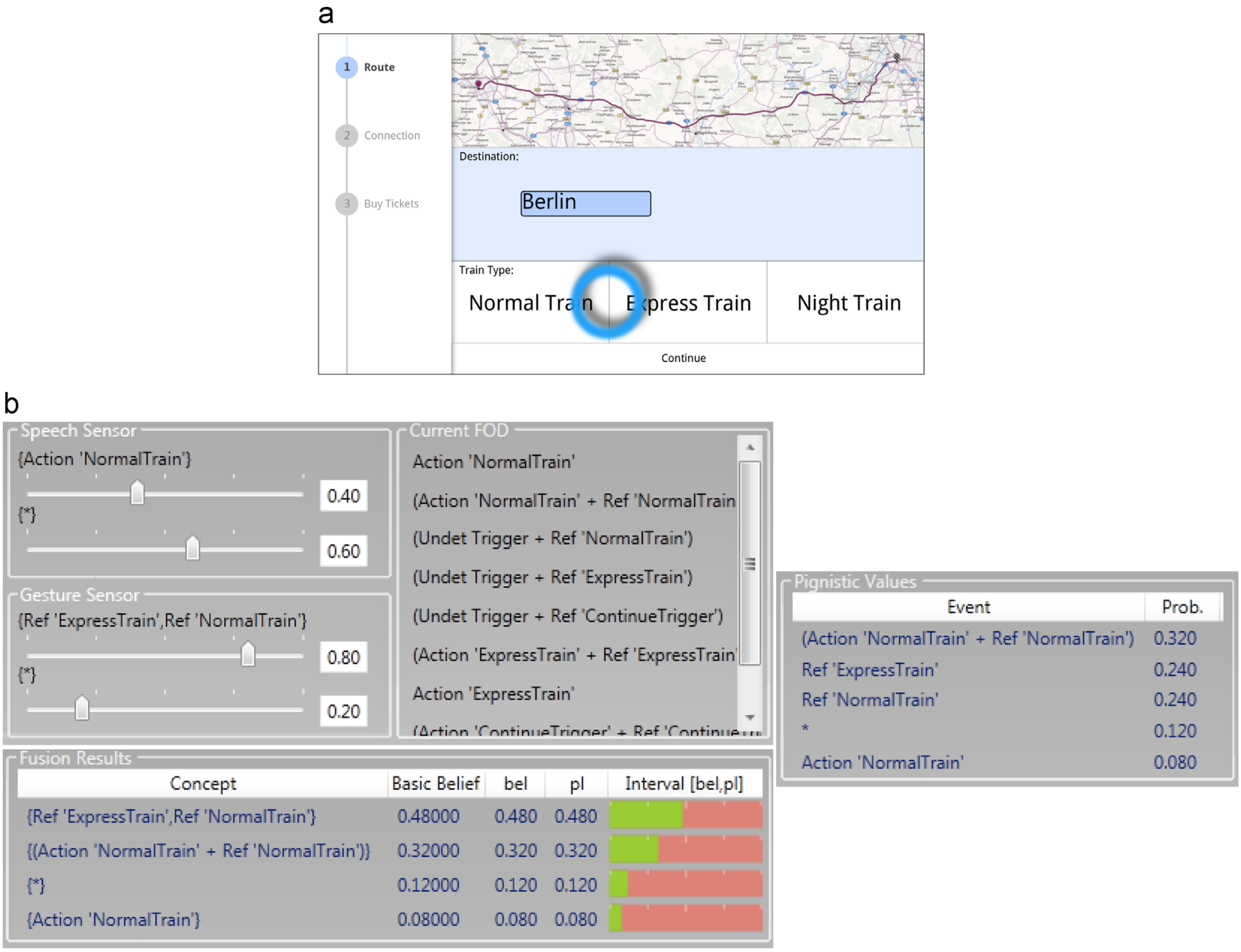
$$m(\mathcal{C}) = \sum_{\mathcal{C}:\mathcal{C} = (\mathcal{A}\times\mathcal{B})\cap\Omega} m_1(\mathcal{A}) \cdot m_2(\mathcal{B}), \tag{18}$$

where $m(\mathcal{X})$ denotes a belief on the set of events $\mathcal{X}$. Given the above example, $\mathcal{A}$ and $\mathcal{B}$ are two sets that contain events a and b respectively. $(\mathcal{A} \times \mathcal{B})$ is the Cartesian product of both sets containing all possible tuples, amongst others it contains (a,b). $(\mathcal{A} \times \mathcal{B}) \cap \Omega$ finally allows only those tuples, out of the Cartesian product $\mathcal{A} \times \mathcal{B}$, that are defined in the FOD.

Once a fusion step triggered by new events is complete, the resulting beliefs are transformed back into probabilities by a pignistic transformation in order to make the final decision on what actually happened. A more detailed explanation of the formal background of the approach is given in [130] and examples for disambiguation, reinforcement and conflict detection are given. An example is given in Fig. 12. Fig. 12a shows a scene from a train ticket booking application, where the user is supposed to select a train type after the destination has already been assigned. In the given situation, the user performs an ambiguous pointing gesture while vaguely uttering “normal train”. Fig. 12b depicts the fusion engine which is visualizing the different inputs and the final result, where inputs reinforce each other.

In the upper part, a low confidence of 0.4 for the action ‘NormalTrain’ (from the speech sensor, where ‘⁎’ denotes an unknown utterance) and ambiguous references for ‘ExpressTrain’ and ‘NormalTrain’ (from the gesture sensor) with a confidence of 0.8 are combined. In the middle part (fusion results), the results of the belief combination according to Eq. (18) are shown. After the pignistic transformation, the highest probability is assigned to the combined event of selecting the ‘NormalTrain’ (pignistic values in the lower part).

To transfer this model to a new application, a FOD stating all possible events has to be defined. In [130] a generic model of events is proposed and a FOD is constructed from a generic description of the current dialog state. While the generic model can be used for some applications, the vast amount of diverse applications of a CTS can hardly ever be covered by a single model, or would require cumbersome tweaks. Thus, we are currently working on enabling arbitrary interaction models that suit the domain at hand using an abstract



Destination: Berlin

Train Type:

Speech Sensor

{Action 'NormalTrain'} 0.40

{*} 0.60

Gesture Sensor

{Ref 'ExpressTrain',Ref 'NormalTrain'} 0.80

{*} 0.20

Current FOD

Action 'NormalTrain'
(Action 'NormalTrain' + Ref 'NormalTrain
(Undet Trigger + Ref 'NormalTrain')
(Undet Trigger + Ref 'ExpressTrain')
(Undet Trigger + Ref 'ContinueTrigger')
(Action 'ExpressTrain' + Ref 'ExpressTrain'
Action 'ExpressTrain'
(Action 'ContinueTrigger' + Ref 'Continue...

Pignistic Values

Event	Prob.
(Action 'NormalTrain' + Ref 'NormalTrain')	0.320
Ref 'ExpressTrain'	0.240
Ref 'NormalTrain'	0.240
*	0.120
Action 'NormalTrain'	0.080

Fusion Results

Concept	Basic Belief	bel	pl	Interval [bel,pl]
{Ref 'ExpressTrain',Ref 'NormalTrain'}	0.48000	0.480	0.480	🟩🟥
{(Action 'NormalTrain' + Ref 'NormalTrain')}	0.32000	0.320	0.320	🟩🟥
{*}	0.12000	0.120	0.120	🟩🟥
{Action 'NormalTrain'}	0.08000	0.080	0.080	🟩🟥

Fig. 12. *Screenshots from a train ticket booking application for input fusion.* (a) Screen capture of the application in which the user is supposed to select a train type. (b) The underlying fusion engine visualizing the inputs. (a) Train ticket booking application user interface. (b) Three stages of the internal processing of beliefs.

graph notation based on GraphML [131]. The graph contains nodes (single events) and edges (combined events) in order to define a FOD. Edges additionally contain XSLT transformations to specify the results of combined events. This allows specifying the semantics of fusion of arbitrary interactions via XML.

Any best scoring result of the multimodal input fusion is passed on to the dialog management, be it a definite input, a still ambiguous input, or even a conflicting input. The dialog management decides on the next step what should be performed in the overall dialog. This topic is discussed in the next section.

4.3. Constraint-based hierarchical dialog management

The task of the dialog management is to coordinate the interaction and dialog progress between human and computer and deciding upon the appropriate system reaction on a given input. Dialog management systems are usually only reactive, i.e. triggered by explicit user input. However, recent trends show an increasing interest in pro-active interactions. The dialog manager may initiate dialogs by itself, triggered by information based on inferences of multiple explicit as well as implicit input sources, cf. Fig. 13.

Commonly used information sources are (1) the current state of the dialog, (2) the dialog history between human and computer, (3) the modeled domain of tasks, and (4) a user model containing characteristics like the user's knowledge. These information sources have to be inferred by the fusion of explicit (e.g. speech, touch, gesture) and implicit (e.g. emotion recognition by physiological signals) user inputs. The decision how to react in the probably most appropriate way to all the present inputs requires the dialog management to perform inference on a large number of essential information from the different available sources. In this context the word “essential” means selecting those parts of information, which are most influential to select the next step in the dialog, and therefore coordinating the dialog flow. The dialog in traditional dialog systems appears to be limited in possibilities of interaction and system functionalities. In contrast, pro-active dialogs appear to be more natural though they are still in large parts event-based. However, the paradigm of pro-active dialog management additionally requires a continuous analysis, inference and evolution of information sources to sense not only the appropriate reaction but also the appropriate time of intervention.

The proposed dialog management, as shown in Fig. 13, is based on the dialog model GEEDI [132] and an explanation manager [133]. The dialog model is a hierarchical tree-like structure consisting of the so-called goals. Goals represent crucial steps in the dialog between human and computer. Every goal has several guards, which are preconditions to be fulfilled in order to accomplish a goal. The guards

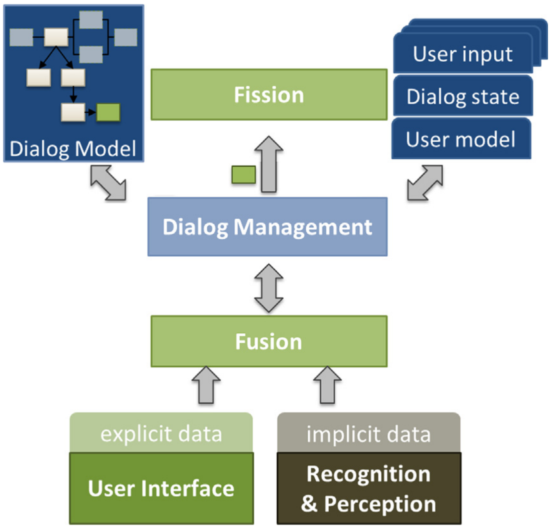


Fig. 13. *The decision process of the dialog management.* Explicit data (e.g. user commands inferred from speech) and implicit data (e.g. affective user states) are combined by the fusion and passed to the dialog management, in order to update the user model and user input history. In the next step, the dialog goal in the dialog model is selected, by taking into account the user input history, the current dialog state and the user model. Only appropriate dialog goals, which satisfy the preconditions (e.g. affective state), are rendered to the fission to be presented to the user.

can monitor many aspects of the dialog such as the emotional state or the state of the user's knowledge. Our approach manages preconditions based on constraint programming [134].

Constraint programming has proven to be particularly well-suited for problems on finite domains where many conditions limit the possible variable value configurations [135]. We consider the preconditions in the guards as constraints. Based on the current values of the variables, deducted from the available information sources, the constraint solver infers which conditions can be fulfilled. The procedure ensures that only those dialogs will be presented to the user that fit the currently available information [136].

The most appropriate dialog goal is selected and transformed into a format that the interaction management's fission component is capable to process. The modalities to convey the content of the dialog goal are determined by the fission component, which will be described in more detail in the next section.

4.4. Fission for modality arbitration

The accessible knowledge derived by multiple fusion steps can be used to support a CTS in different ways. Our goal is to make use of acquired knowledge which describes a certain context of use (CoU) in order to provide the user with an individual system behavior as well as to tailor the offered functionality. Such an individual behavior does not only affect the adaptive planning and dialog management components, but does also apply to the process of fission, which is in charge to reason about the individual user interface.

Ideally, a CTS adapts its functionality and user interface to the particular user according to the actual CoU. To realize an individual user interface, the system has to consider the abilities of the user, his preferences, as well as optional requirements that can arise in the CoU. Hence, continual perception and adaption is the key to realize a system being permanently available, cooperative, and trustworthy.

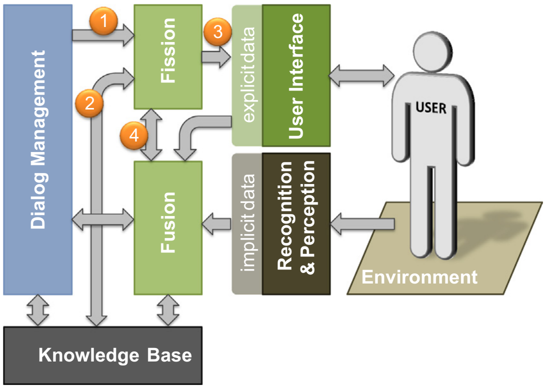


Fig. 14. *The fission and the linked components.* (1) A modality-independent output is passed to the fission. The fission inspects the current knowledge, (2) and reasons about the proper output representation for the detected context of use. (3) The modality-specific output configuration is passed to the rendering user interface components. (4) The fission interacts with the fusion component to be able to adapt to the user's preferences.

The procedure of modality arbitration is the main part of the multimodal fission process. The interplay with the other components is depicted in Fig. 14. The term “adaptive modality arbitration” means that in a first instance the system, e.g. the dialog management, describes the output in an abstract and modality-independent manner. This description is then passed to the fission for further processing (cf. Fig. 14, ①). For example, the dialog management initiates the output of a train time table. At that point of time, only the abstract output of a certain information item, the *time table*, is announced. The decision about which widget will render the time table on which device, or which combination of modalities shall be used to present the output to the user, is dedicated to the fission module. The fission module reasons about the adequate, and final output configuration [124,125]. Rousseau et al. [125] state that the main tasks in fission are concerned with the following four WWHT-questions: (1) “*What* is the information to present?” (2) “*Which* modalities should be used to present this information?” (3) “*How* to present the information using these modalities?” (4) and “*Then*, how to handle the evolution of the resulting presentation?”.

The first question is directly answered by the abstract information provided by the dialog management (cf. Fig. 14, ①). Regarding the second and the third question, the fission module interacts with the knowledge base (KB) to fetch possible mappings for each communicable information fragment from its abstract to its concrete form of representation (cf. Fig. 14, ②).

In the case of the train time table, this could be a textual, an aural, or a pictorial representation. Furthermore, the fission module queries the device models from the KB, which are describing the present interaction hardware with its possibilities for rendering and interaction. This data allows us to answer the *what* and *how* questions. Besides the user-independent information, the KB additionally provides recent user and environment models (cf. Fig. 14, ②) which describe the user in his current CoU. Due to the real-world setting, the proposed approach models uncertain statements using a probability distribution.

Next, the fission module identifies all possible output configurations in their unimodal and combinational multimodal manifold in order to reason about the most adequate output configuration for the current CoU. The current CoU is described by the set $\mathcal{K}$, which contains mutually independent variables K, e.g. “gender” or “environmental noise level”. These variables are composed of basic elements k, i.e. in case of the variable gender the corresponding elements are female and male. The perception of the CTS assigns probabilities to the basic elements ψ_k. Possible

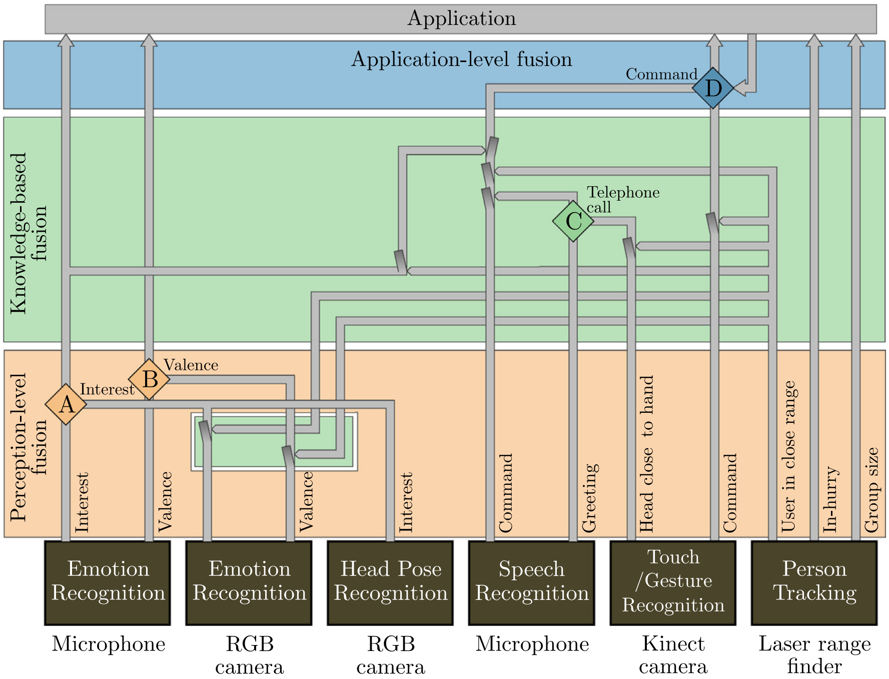


Fig. 15. *Overview of the implemented integrated system.*The output of the recognition components is combined on different levels. Important parts of the system are labeled with A, B, C, D. For a detailed description refer to the text.

output configurations oc are weighted by a set of evaluation functions $f \in \mathcal{F}^{oc}$. Each function f represents a pre-defined rule from a given design and style guide and focuses on a subset of M^f basic elements denoted by $\{k^f_1, \dots, k^f_{M^f}\}$. The function f is defined as a weighted product of probabilities:

$$f(\psi^f_{k_1}, \dots, \psi^f_{k_{M^f}}) = w^f \cdot \prod_{m=1}^{M^f} \psi_{k^f_m} \tag{19}$$

where $w^f \in [-1, 1]$ is an individual weight provided for each function. The highest-rated output configuration $oc^\star$ represents the final candidate:

$$oc^\star = \max_{oc} \sum_{f \in \mathcal{F}^{oc}} f(\psi^f_{k_1}, \dots, \psi^f_{k_{M^f}}) \tag{20}$$

which is then passed to the rendering UI device components (cf. Fig. 14, ③). In addition, the knowledge base and the fusion component are updated with the current output configuration to adjust and adapt their ongoing reasoning (cf. Fig. 14, ② and ④).

The forth question, concerning the temporal evolution of the UI, is solved by a continuous reasoning process triggered by any influencing factor within the observed CoU. For instance, if the user has changed its position, e.g. as described in Section 3.3, another output component might be more suitable to render the current information fragments as part of the dialog. If required, our current implementation can initiate a new fission process every 500 ms. The whole process of modality arbitration is described in detail in [10]. Different context models and different levels of adaption are described in [137]. A more detailed interplay of the two components for multimodal fusion and fission is described in [138].

5. Implementation

The approaches presented in the previous sections show that the tasks to be handled in a CTS are computationally expensive, simply because of the large number of components processing simultaneously. However, since the information has to be combined gradually, it is possible to effectively distribute the load to a compound of processing units. This section presents a working example of a distributed multi-layered fusion architecture following a uniform design principle. Fig. 16 shows the realization of the hardware platform. This platform is constructed for a general HCI use such that various applications can be implemented [139,140]. For the sake of an example the platform serves as a cognitive ticket vending application throughout this section. The setting requires a set of input channels and output channels to be connected to the system. The sensory input channels comprise a high-resolution camera,[1] a stereo-camera,[2] a Kinect camera, a touch display, two laser range finders,[3] and a head-set microphone. The output is realized by a display and a sound system. Additional screens visualize the latest information about the system for further analysis, i.e. the current internal state with details of specific system components.

The tracking algorithm operating on the laser range finders determines the size of the group from which the main user separates and the main user's speed while approaching the system. The location of the main user is of high importance for the algorithms' working on the data of the attached cameras: Only faces or bodies, which are located at the estimated position of the main user, are allowed for further processing. The data of the Kinect camera is used to realize a gesture recognition, where the gestures are interpreted as explicit commands to the system. The stereo camera is utilized to perform a head pose recognition. An orientation towards the system is interpreted as an “interest” for interaction. The high-resolution camera is used to perform a facial emotion recognition on the categories “valence” and “interest”. The speech input is used to capture commands and implicit non-verbal audio cues from the same categories as for visual emotion recognition. All recognizers operate on their maximal frame rates and with the fastest response time possible.

[1] Pike from Allied Vision.

[2] Bumblebee from Point Grey.

[3] Ibeo Lux from Ibeo Automotive Systems.

The basic recognition results are used to perform fusion on different layers according to the proposed schema. A schematic drawing providing an overview of the fusion layers is shown in Fig. 15. At the bottom of the figure, different sensor channels are listed which serve as input for the recognition components. The outputs of the recognition components are visualized using gray arrows leading towards the application. The fusion on the perception-level is performed as described in Section 2.4, and requires all inputs to be from the same class. Label A marks the combination of the class “interest” using the output from the head pose analysis plus the aural and visual emotion recognition. Knowledge from higher level is used to gate results of the facial emotion recognition, and to allow only measurements derived from the main user. The corresponding region is highlighted by a green area in the layer of perceptive-level fusion. Label B marks the combination of the class “arousal” using only the auditory and visual emotion recognition. Therefore, a set of rules has been applied in the knowledge-based fusion layer. The most important rule ensures that image-based recognitions are truly originated from the main user. One exemplary rule, marked by the label C, denotes a multimodal fusion rule to recognize a telephone call. On application-level, the multimodal input fusion, described in Section 4.2, is performed. The fusion is marked on the figure by the label D and retrieves additional data from the application as indicated by an arrow leading from the application to the fusion.

In the ticket vending scenario, the personal knowledge about the user includes the user's ticket preference. For instance, he might prefer faster connections or fewer changes of trains. The vending machine could also make use of the user's personal appointment calendar. In order to acquire that kind of information, the application has to query a personal data service to provide the knowledge base with this additional data.

Based on diverse sensors, the system can be controlled via gesture, touch, or speech commands. It pro-actively opens the dialog when a subject approaches. Furthermore, the dialog's content (e.g. the number of tickets) or the complexity of the user interface is adapted based on the recognized size of the group associated with the user and the recognized level of hurry, which is determined by the approaching speed of the user. In crucial moments of interaction, e.g. the dialog guesses the group size, the system interprets the emotional state of the user in order to ensure that the user is still pleased with the progress of the dialog and the system's decisions. In case the system detects a phone conversation (as in Fig. 16) or a lowered interest, the machine will no longer follow the commands received via speech input. Similarly, the components or emotion and gesture recognition are deactivated in cases where the user turns away.

Each algorithmic approach, including the fusion (but except the touch input which is co-located with the algorithms in the application layer), is placed on a separated processing unit. The emotion, head pose, and body pose recognitions are based on different video channels, which are operating at 15 Hz, whereas the tracking of persons can provide predictions at a frequency of 100 Hz. The gesture and speech recognition modules provide information in an event-based manner. The information exchange does not only happen in one direction, but often requires a feedback to other components in order to reduce the computational effort, e.g. to deactivate the emotion recognition in case the user shifts attention or to suppress the speech recognition in order to respect the user's privacy.

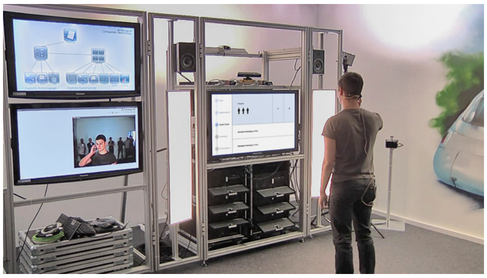


Fig. 16. *The realized prototypical system.* The two screens on the left allow an insight into the system's processing. The screen on the right supports the actual interaction.

A second prototype using the same hardware configuration, but with a complementary focus on the processing of symbolic information, has been described in [139]. The application is designed to individually assist the user to assemble a home theater system and makes use of the three approaches presented in Section 4, the input fusion, the constraint-based dialog management plus the fission for modality arbitration. Furthermore, the application uses a planning component, which is able to adapt to the actual situation by following a plan repair approach.

The CTS has been operated using six retail computers with the following resource allocation: (1) Semaine server, speech recognition, application; (2) auditory emotion recognition, Markov fusion networks, activity recognition; (3) facial emotion recognition; (4) gesture recognition; (5) person tracking; (6) head pose recognition. Since the algorithms and system load have not been optimized for efficiency, a smaller set of computers will be sufficient for the task at hand. The actual computational demand will become more and more insignificant in the light of the ubiquitous smart home devices and the“internet of things” development There are different kinds of middleware architectures which have meet our requirements for both prototypes, and which would allow the desired form of communication of our components [141–143]. For the both CTS we decided upon a message-based middleware, namely the Java Messaging Service (JMS), since this type of architecture allows both for a peer-to-peer connection and a broadcast communication. The wrapper OpenSource framework of the Semaine-Project [144] is utilized as a common basis to interface the JMS and has been extended by an iOS integration using a Semaine proxy client, a C client and a MQTT wrapper to support mobile devices.

6. Discussion

We introduced an alternative view to the field of information fusion. Classical fusion approaches usually combine data derived almost directly from the sensors and are often tied to a single recognition task with a set of identical classes. This work proposes a categorization for CTS into three fusion levels, namely perception fusion, knowledge-based fusion and application-level fusion.

The perception fusion represents the natural foundation of a CTS since the recognizers located in this layer are the first ones processing the data from the sensory. Hence, this layer has to provide robust and continuous recognitions of the user's state and his environment. In Section 2, we presented the use of dynamic features and novel fusion strategies for affective state recognition. The well-known Kalman filter and the MFN have been introduced to handle the absence of classifier decisions in a continuous stream of decisions, e.g. due to rejection or sensor failures. The results show that the novel probabilistic temporal fusion approaches clearly outperform the uni-modal results, while at the same time providing a mechanism to compensate missing classifier decisions. Compared to the probabilistic temporal fusion approaches, the adaptive fusion of dynamic features, introduced in Section 2.2, bears the advantage that temporal patterns can be learned with the help of a time window. The MFN and the Kalman filter for classifier fusion both are applied in the same usage scenarios. The Kalman filter explicitly models the uncertainty of classifier measurements, whereas the MFN excels through a large range of options for parameterization. In practice, the performance of both probabilistic temporal fusion approaches can be regarded as very good. The CTS implementation makes use of the MFN to combine the results of the classifiers operating on multiple modalities, recognizing the user state "interest" and the affective state "valence". The successful application of the MFN shows that the algorithm meets the requirement of real-time scenarios and provides an example of how affective recognition results can support the dialog management in critical moments.

The next layer, named knowledge-based fusion, builds up on the classes recognized in the former level and intends to recognize more complex classes with the help of additional context information and observations from larger time scales. The identification of complex classes generally requires the recognition of intermediate classes which need to be combined to derive the final class [108]. The approaches presented in this work combine information derived from sub-symbolic recognitions with additional generic symbolic rules. Section 3 provides three approaches: (1) the modeling of sensor properties in person tracking using DST, (2) the association of locations based on a first-order probabilistic model, and (3) the layered application of Markov models. The first approach enhances multi-object tracking by exploiting the different characteristics of two modalities in order to ingeniously combine a set of disjunct classes. It shows how prior knowledge of the modalities can be incorporated by using Dempster's rule of combination. Results on classical approaches are clearly outperformed by this new technique. The second approach utilizes first-order logic to obtain the correct association of tracking measurements and persons in a familiar environment. The idea is to allow obstacles and personal goals to influence the association process. It was shown that the number of false associations can be significantly reduced in comparison to classical approaches. The knowledge-based fusion approach makes use of a layered architecture to recognize classes with increasing complexity. Three different MLN have been evaluated to show the advantages of using additional knowledge. The three approaches document the fact that there is a wide range of fusion applications with a variety of techniques to be studied in the near future. Core idea is to closely integrate the symbolic and the sub-symbolic processing with a minimal loss of information to perform a mapping to more abstract class sets. It has been shown that it is beneficial to make use of class confidence measures. Furthermore, it is important to mind the presence of relational data, e.g. spatial, temporal. The implemented prototype used the knowledge-based fusion in many ways, e.g. to recognize an ongoing telephone call. However, the availability of a large set of recognition results on different sensory inputs in a CTS makes it seem natural to exploit their synergies. A good example is suppressing the affective video recognition results in case the user is not facing the camera. Due to the large variety of applications and settings, it is important to provide common datasets in this area to ensure a constructive research.

The final CTS fusion layer is called the application-level fusion. The algorithms in this category mediate between information provided by the knowledge base, the application, and the explicit and implicit user inputs. The first presented approach infers the user's input by combining multiple input modalities and the options provided by the application using the DST. The second approach focuses on the dialog management which has to provide an efficient dialog based on all available information, i.e. implicit and explicit user inputs and the overall communication context. The last approach enhances the HCI by analyzing the information about the user, e.g. position or interaction modality, in order to predict the most appropriate output modality. All three presented approaches follow different strategies: (1) applying a probabilistic calculus (3) using constraint programming, and (3) rule-based weighting. In doing so, the algorithms predict the user's goal, combine multiple uncertain sources of information and find the best crisp decision. Likewise to the knowledge-based fusion, the explicit interface realized by application-level fusion bears a large variety of applications since a large number of consistent decisions have to be made in a CTS. The explicit user input of the implemented example is obtained using the first approach. To do so, the belief input fusion mediates between touch, speech, gestural input and the available options provided by the system. The two other presented approaches have not been realized in this setting since the share of HCI was not been rich enough. The application-level fusion enables the user to have a natural discourse with the CTS. It shows the full potential in settings with a wide range of available options. For instance, in a setting in which a companion system becomes part of the user's everyday life such that the CTS can build up a sophisticated user model.

The implemented prototype takes advantage of multimodal input and output channels to allow a natural interaction with the system. However, handling such a large number of diverse channels requires a carefully thought out processing in order to make full use of the synergy effects. First tendencies of realizing such a close interplay between modalities can already be found in retail products. However, on closer look it becomes evident that the larger context has usually a subordinate role and only a simple control circuit is realized, e.g. the movie player stops in case the spectator turns away. The implemented CTS uses the collected information in a different manner, e.g. with a larger time span (the number of traveler) or as part of an adaptive dialog (interest or emotional valence). Only a subset of the presented approaches have been integrated into the final implementation since the development of a complete system would have bound to many human resources. However, the example shows that the expressiveness of a system grows with the amount of merged information and, as a result, offers new possibilities for powerful applications which can put special focus on features such as adaptivity, individualization or cooperativity.

7. Conclusions

The present paper gives an overview on information fusion challenges in CTS and provides an outlook to algorithmic demands in the future. We showed that important features (among others) are the fusion of multiple sources over time, the seamless end-to-end handling of uncertainty and the back-propagation of high-level information derived within the inner building block of the CTS. We categorized the information fusion approaches present in a CTS into three categories: (1) processing of implicitly given information; (2) explicit input and output, which have to be consistent with both the application and the user intentions; and (3) the integration of high-level information itself. The new taxonomy has been exemplified by nine algorithmic approaches and validated by a representative implementation, which unites components introduced previously. The prototypical realization of CTS does not only show that state-of-the-art approaches are already qualified to build first instances of such CTS, but also that studies have to increasingly focus on the interplay of different research fields, e.g. pattern recognition plus artificial intelligence. The information exchange and fusion within and between the proposed categories may lead to synergies bringing decisive improvements compared to single approaches. Future work will aim at recording and studying an interaction scenario placing special emphasis on the principle of fusion and fission of information.

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).

References

[1] A. Wendemuth, S. Biundo, A companion technology for cognitive technical systems, in: A. Esposito, A.M. Esposito, A. Vinciarelli, R. Hoffmann, V.C. Muller (Eds.), Cognitive Behavioural Systems, Lecture Notes in Computer Science, vol. 7403, Springer, Berlin, Heidelberg, 2012, pp. 89–103. http://dx.doi.org/10.1007/978-3-642-34584-5_7.

[2] G. Palm, M. Glodek, Towards emotion recognition in human computer interaction, in: A. Esposito, S. Squartini, G. Palm (Eds.), Neural Nets and Surroundings, Smart Innovation Systems and Technologies, vol. 19, Springer, Berlin, Heidelberg, 2013, pp. 323–336. http://dx.doi.org/10.1007/978-3-642-35467-0_32.

[3] A. Schmidt, Implicit human computer interaction through context, Pers. Technol. 4 (2–3) (2000) 191–199. http://dx.doi.org/10.1007/BF01324126.

[4] S. Scherer, M. Glodek, G. Layher, M. Schels, M. Schmidt, T. Brosch, S. Tschechne, F. Schwenker, H. Neumann, G. Palm, 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) (2012) 117–141. http://dx.doi.org/10.1007/s12193-012-0093-9.

[5] A.K. Dey, G.D. Abowd, Towards a better understanding of context and context-awareness, in: H.-W. Gellersen (Ed.), Proceedings of the International Symposium on Handheld and Ubiquitous Computing, Lecture Notes in Computer Science, vol. 1707, Springer, London, UK, 1999, pp. 304–307. http://dx.doi.org/10.1007/3-540-48157-5_29.

[6] M. Glodek, M. Schels, G. Palm, F. Schwenker, Multi-modal fusion based on classifiers using reject options and Markov fusion networks, in: Proceedings of the International Conference on Pattern Recognition (ICPR), IEEE, 2012, pp. 1084–1087.

[7] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2009, pp. 365–372. http://dx.doi.org/10.1109/ICCV.2009.5459250.

[8] M. Glodek, G. Layher, F. Schwenker, G. Palm, Recognizing human activities using a layered Markov architecture, in: A.E. Villa, W. Duch, P. Érdi, F. Masulli, G. Palm (Eds.), Proceedings of the International Conference on Artificial Neural Networks and Machine Learning (ICANN), Lecture Notes in Computer Science, vol. 7552, Springer, Berlin, Heidelberg, 2012, pp. 677–684. http://dx.doi.org/10.1007/978-3-642-33269-2_85.

[9] T. Geier, S. Reuter, K. Dietmayer, S. Biundo, Track-person association using a first-order probabilistic model, in: Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI), vol. 1, IEEE, 2012, pp. 844–851. http://dx.doi.org/10.1109/ICTAI.2012.118.

[10] F. Honold, F. Schüssel, M. Weber, Adaptive probabilistic fission for multi-modal systems, in: Proceedings of the Australian Computer–Human Interaction Conference (OzCHI), ACM, New York, NY, 2012, pp. 222–231. http://dx.doi.org/10.1145/2414536.2414575.

[11] M. Buss, M. Beetz, CoTeSys—Cognition for technical systems, Kunstl. Intell. 24 (4) (2010) 323–327. http://dx.doi.org/10.1007/s13218-010-0061-z.

[12] H.P. Moravec, The stanford cart and the CMU rover, Proc. IEEE 71 (7) (1983) 872–884. http://dx.doi.org/10.1109/PROC.1983.12684.

[13] J. Laird, A. Newell, P. Rosenbloom, SOAR: an architecture for general intelligence, Artif. Intell. 33 (1) (1987) 1–64. http://dx.doi.org/10.1016/0004-3702(87)90050-6.

[14] J.R. Anderson, M. Matessa, C. Lebiere, A theory of higher level cognition and its relation to visual attention, Hum. Comput. Interact. 12 (4) (1997) 439–462. http://dx.doi.org/10.1207/s15327051hci1204_5.

[15] R. Sun, A Tutorial on CLARION 5.0, Cognitive Science Department, Rensselaer Polytechnic Institute, URL ⟨http://www.cogsci.rpi.edu/∼rsun/clarion.html⟩, 2003 (last visited 01/10/2013).

[16] A. Newell, Unified Theories of Cognition, Harvard University Press, 1994.

[17] J.R. Anderson, C. Lebiere, The Newell test for a theory of cognition, Behav. Brain Sci. 26 (5) (2003) 587–639. http://dx.doi.org/10.1017/S0140525X0300013X.

[18] D. Vernon, G. Metta, G. Sandini, A survey of artificial cognitive systems: implications for the autonomous development of mental capabilities in computational agents, IEEE Trans. Evol. Comput. 11 (2) (2007) 151–180. http://dx.doi.org/10.1109/TEVC.2006.890274.

[19] G.H. Granlund, The complexity of vision, Signal Process. 74 (1) (1999) 101–126. http://dx.doi.org/10.1016/S0165-1684(98)00204-7.

[20] Y. Mohammad, T. Nishida, Controlling gaze with an embodied interactive control architecture, Appl. Intell. 32 (2) (2010) 148–163. http://dx.doi.org/10.1007/s10489-009-0180-0.

[21] J.R. Anderson, ACT: a simple theory of complex cognition, Am. Psychol. 51 (4) (1996) 355–365. http://dx.doi.org/10.1037/0003-066X.51.4.355.

[22] J. Ball, Explorations in ACT-R based language analysis—memory chunk activation, in: N. Rußwinkel, U. Drewitz, H. van Rijn (Eds.), Proceedings of the International Conference on Cognitive Modeling, Universitätsverlag der TU, Berlin, 2012, pp. 131–136.

[23] G. Fink, N. Jungclaus, F. Kummert, H. Ritter, G. Sagerer, A distributed system for integrated speech and image understanding, in: Proceedings of the International Symposium on Artificial Intelligence (ISAI/IFIS) Collaboration in Intelligent Systems Technologies, 1996, pp. 117–126.

[24] N. Jungclaus, M.v.d. Heyde, H. Ritter, G. Sagerer, An architecture for distributed visual memory, Z. Naturforschung C (A Journal of Biosciences) 53 (7/8) (1998) 550–559.

[25] W. Wahlster, SmartKom: Symmetric multimodality in an adaptive and reusable dialogue shell, in: R. Krahl, D. Günther (Eds.), Proceedings of the Status Conference “Human Computer Interaction”, DLR, 2003, pp. 47–62.

[26] G. Herzog, N. Reithinger, The SmartKom architecture: a framework for multimodal dialogue systems, in: W. Wahlster (Ed.), SmartKom: Foundations of Multimodal Dialogue Systems, Cognitive Technologies, Springer, Berlin, Heidelberg, 2006, pp. 55–70. http://dx.doi.org/10.1007/3-540-36678-4_4.

[27] C. Burghart, R. Mikut, R. Stiefelhagen, T. Asfour, H. Holzapfel, P. Steinhaus, R. Dillmann, A cognitive architecture for a humanoid robot: a first approach, in: Proceedings of the IEEE-RAS International Conference on Humanoid Robots, IEEE, 2005, pp. 357–362. http://dx.doi.org/10.1109/ICHR.2005.1573593.

[28] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, Inc., 2004 http://dx.doi.org/10.1002/0471660264.

[29] C. Dietrich, G. Palm, K. Riede, F. Schwenker, Classification of bioacoustic time series based on the combination of global and local decisions, Pattern Recognit. 37 (12) (2004) 2293–2305. http://dx.doi.org/10.1016/j.patcog.2004.04.004.

[30] C. Dietrich, G. Palm, F. Schwenker, Decision templates for the classification of bioacoustic time series, Inf. Fusion 3 (2) (2003) 101–109. http://dx.doi.org/10.1016/S1566-2535(03)00017-4.

[31] H. Wallbott, K. Scherer, Cues and channels in emotion recognition, J. Personal. Social Psychol. 51 (4) (1986) 690–699.

[32] A. Vinciarelli, M. Pantic, H. Bourlard, A. Pentland, 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, New York, NY2008, pp. 1061–1070. http://dx.doi.org/10.1145/1459359.1459573.

[33] S. Scherer, M. Glodek, F. Schwenker, N. Campbell, G. Palm, Spotting laughter in natural multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data, ACM Trans. Interact. Intell. Syst. (Special Issue on Affective Interaction in Natural Environments) 2 (1) (2012) 4:1–4:31. http://dx.doi.org/10.1145/2133366.2133370.

[34] A. Panning, I. Siegert, A. Al-Hamadi, A. Wendemuth, D. Rösner, J. Frommer, G. Krell, B. Michaelis, Multimodal affect recognition in spontaneous HCI environment, in: Proceedings of the IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC), ACM , New York, NY, 2012, pp. 430–435. http://dx.doi.org/10.1109/ICSPCC.2012.6335662.

[35] G. Krell, M. Glodek, A. Panning, I. Siegert, B. Michaelis, A. Wendemuth, F. Schwenker, Fusion of fragmentary classifier decisions for affective state recognition, in: F. Schwenker, S. Scherer, L.-P. Morency (Eds.), Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Lecture Notes in Computer Science, vol. 7742, Springer, Berlin, Heidelberg, 2012, pp. 116–130. http://dx.doi.org/10.1007/978-3-642-37081-6_13.

[36] P. Ekman, W.V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978.

[37] P. Ekman, Facial expression and emotion, Am. Psychol. 48 (4) (1993) 384–392. http://dx.doi.org/10.1037/0003-066X.48.4.384.

[38] R. Niese, A. Al-Hamadi, M. Heuer, B. Michaelis, B. Matuszewski, Machine vision based recognition of emotions using the circumplex model of affect, in: Proceedings of the International Conference on Multimedia Technology (ICMT), IEEE, 2011, pp. 6424–6427. http://dx.doi.org/10.1109/ICMT.2011.6001887.

[39] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, M. Bartlett, The computer expression recognition toolbox (CERT), in: Proceedings of the International Conference IEEE on Automatic Face Gesture Recognition and Workshops (FG), IEEE, 2011, pp. 298–305. http://dx.doi.org/10.1109/FG.2011.5771414.

[40] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, Emotional speech: towards a new generation of databases, Speech Commun. 40 (1–2) (2003) 33–60. http://dx.doi.org/10.1016/S0167-6393(02)00070-5.

[41] M. Glodek, S. Reuter, M. Schels, K. Dietmayer, F. Schwenker, Kalman filter based classifier fusion for affective state recognition, in: Z.-H. Zhou, F. Roli, J. Kittler (Eds.), Multiple Classifier Systems (MCS), Lecture Notes in Computer Science, vol. 7872, Springer, Berlin, Heidelberg, 2013, pp. 85–94. http://dx.doi.org/10.1007/978-3-642-38067-9_8.

[42] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140. http://dx.doi.org/10.1007/BF00058655.

[43] M. Glodek, M. Schels, F. Schwenker, Ensemble Gaussian mixture models for probability density estimation, Comput. Stat. 28 (1) (2013) 127–138. http://dx.doi.org/10.1007/s00180-012-0374-5.

[44] N. Dahlbäck, A. Jönsson, L. Ahrenberg, Wizard of Oz studies—why and how, Knowl. Based Syst. 6 (4) (1993) 258–266. http://dx.doi.org/10.1016/0950-7051(93)90017-N, special issue: Intelligent User Interfaces.

[45] D. Rösner, J. Frommer, R. Friesen, M. Haase, J. Lange, M. Otto, LAST MINUTE: a multimodal corpus of speech-based user-companion interactions, in: N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the International Conference on Language Resources and Evaluation Conference (LREC), European Language Resources Association (ELRA), 2012, pp. 23–25.

[46] D.O. North, An analysis of the factors which determine signal/noise discrimination in pulsed-carrier systems, Proc. IEEE 51 (7) (1963) 1016–1027. http://dx.doi.org/10.1109/PROC.1963.2383.

[47] G. Krell, R. Niese, A. Al-Hamadi, B. Michaelis, Suppression of uncertainties at emotional transitions—facial mimics recognition in video with 3-D model, in: P. Richard, J. Braz (Eds.), Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, 2010, pp. 537–542.

[48] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239. http://dx.doi.org/10.1109/34.667881.

[49] D. Ruta, B. Gabrys, An overview of classifier fusion methods, Comput. Inf. Syst. 7 (1) (2000) 1–10.

[50] C. Thiel, Multiple Classifier Systems Incorporating Uncertainty, Verlag Dr. Hut, 2010.

[51] T. Ho, J. Hull, S. Srihari, Decision combination in multiple classifier systems, IEEE Trans. Pattern Anal. Mach. Intell. 16 (1) (1994) 66–75. http://dx.doi.org/10.1109/34.273716.

[52] B. Jeon, D.A. Landgrebe, Decision fusion approach for multitemporal classification, IEEE Trans. Geosci. Remote Sens. 37 (3) (1999) 1227–1233. http://dx.doi.org/10.1109/36.763278.

[53] L.I. Kuncheva, J.C. Bezdek, R.P.W. Duin, Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognit. 34 (2) (2001) 299–314. http://dx.doi.org/10.1016/S0031-3203(99)00223-X.

[54] I. Bloch, A. Hunter, A. Appriou, A. Ayoun, S. Benferhat, P. Besnard, L. Cholvy, R. Cooke, F. Cuppens, D. Dubois, et al., Fusion: general concepts and characteristics, Int. J. Intell. Syst. 16 (10) (2001) 1107–1134. http://dx.doi.org/10.1002/int.1052.

[55] G. Giacinto, F. Roli, Design of multiple classifier systems, in: Hybrid Methods in Pattern Recognition, vol. 47, World Scientific Publishing, 2002, pp. 199–226 (Chapter 8). http://dx.doi.org/10.1142/9789812778147_0008.

[56] C. Sanderson, K.K. Paliwal, Identity verification using speech and face information, Digit. Signal Process. 14 (5) (2004) 449–480.

[57] F. Bach, G. Lanckriet, M. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, in: Proceedings of the International Conference on Machine Learning (ICML), ACM, New York, NY, 2004, pp. 321–327. http://dx.doi.org/10.1145/1015330.1015424.

[58] N. Poh, J. Kittler, Multimodal information fusion, in: Multimodal Signal Processing, Academic Press, 2010, pp. 153–169 (Chapter 8).

[59] F. Schwenker, C.R. Dietrich, C. Thiel, G. Palm, Learning of decision fusion mappings for pattern recognition, J. Artif. Intell. Mach. Learn. (2006) 17–21 (Special issue: Multiple Classifier Systems).

[60] M. Schels, M. Glodek, G. Palm, F. Schwenker, Revisiting AVEC 2011—an information fusion architecture, in: A. Esposito, S. Squartini, G. Palm, B. Apolloni, S. Bassis, A. Esposito, F.C. Morabito (Eds.), Neural Nets and Surroundings, Smart Innovation, Systems and Technologies, vol. 19, Springer, Berlin, Heidelberg, 2013, pp. 385–393. http://dx.doi.org/10.1007/978-3-642-35467-0_38.

[61] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, F. Schwenker, Multiple classifier systems for the classification of audio-visual emotional states, in: S. D'Mello, A. Graesser, B. Schuller, J.-C. Martin (Eds.), Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science, vol. 6975, Springer, Berlin, Heidelberg, 2011, pp. 359–368. http://dx.doi.org/10.1007/978-3-642-24571-8_47.

[62] M. Schels, S. Scherer, M. Glodek, H. Kestler, G. Palm, F. Schwenker, On the discovery of events in EEG data utilizing information fusion, Comput. Stat. 28 (1) (2013) 5–18. http://dx.doi.org/10.1007/s00180-011-0292-y.

[63] M. Schels, M. Glodek, S. Meudt, S. Scherer, M. Schmidt, G. Layher, S. Tschechne, T. Brosch, D. Hrabal, S. Walter, G. Palm, H. Neumann, H. Traue, F. Schwenker, Multi-modal classifier-fusion for the recognition of emotions, in: Coverbal synchrony in Human–Machine Interaction, CRC Press, 2013, pp. 73–97.

[64] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, NY, 2006.

[65] C. Thiel, F. Schwenker, G. Palm, Using Dempster-Shafer theory in MCF systems to reject samples, in: N. Oza, R. Polikar, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 3541, 2005, pp. 118–127. http://dx.doi.org/10.1007/11494683_12.

[66] R.E. Kalman, A new approach to linear filtering and prediction problems, Trans. ASME—J. Basic Eng. 82 (Series D) (1960) 35–45.

[67] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, M. Pantic, AVEC 2011—The first international audio visual emotion challenges, in: S. D'Mello, A. Graesser, B. Schuller, J.-C. Martin (Eds.), Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), Lecture Notes in Computer Science, vol. 6975, Springer, Berlin, Heidelberg, 2011, pp. 415–424. http://dx.doi.org/10.1007/978-3-642-24571-8_53.

[68] G. McKeown, M. Valstar, R. Cowie, M. Pantic, The SEMAINE corpus of emotionally coloured character interactions, in: Proceedings of the International Conference on Multimedia and Expo (ICME), IEEE, 2010, pp. 1079–1084. http://dx.doi.org/10.1109/ICME.2010.5583006.

[69] M. Glodek, M. Schels, G. Palm, F. Schwenker, Multiple classifier combination using reject options and Markov fusion networks, in: Proceedings of the International ACM Conference on Multimodal Interaction (ICMI), ACM, New York, NY, 2012, pp. 465–472. http://dx.doi.org/10.1145/2388676.2388778.

[70] M. Glodek, M. Schels, F. Schwenker, G. Palm, Combination of sequential class distributions from multiple channels using Markov fusion networks, J. Multimodal User Interfaces 8 (3) (2014) 257–272. http://dx.doi.org/10.1007/s12193-014-0149-0.

[71] J. Dinsmore, D.J. Chalmers, F. Adams, K. Aizawa, G. Fuller, J. Schwartz, B. Douglas S, L.A. Meeden, J.B. Marshall, J.A. Barnden, C.-D. Lee, M. Gasser, S.C. Kwasny, K.A. Faisal, T.E. Lange, The Symbolic and Connectionist Paradigms: Closing the Gap, Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 1992.

[72] R. Möller, B. Neumann, Ontology-based reasoning techniques for multimedia interpretation and retrieval, in: Y. Kompatsiaris, P. Hobson (Eds.), Semantic Multimedia and Ontologies, Part II, Springer, London, 2008, pp. 55–98. http://dx.doi.org/10.1007/978-1-84800-076-6_3.

[73] M. Richardson, P. Domingos, Markov logic networks, Mach. Learn. 62 (1–2) (2006) 107–136. http://dx.doi.org/10.1007/s10994-006-5833-1.

[74] F. Müller, C. Späth, T. Geier, S. Biundo, Exploiting expert knowledge in factored POMDPs, in: L.D. Raedt, C. Bessière, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, P.J.F. Lucas (Eds.), Proceedings of the European Conference on Artificial Intelligence (ECAI), vol. 242, IOS Press, 2012, pp. 606–611. http://dx.doi.org/10.3233/978-1-61499-098-7-606.

[75] G. Shafer, The Dempster-Shafer theory, in: S.C. Shapiro (Ed.), Encyclopedia of Artificial Intelligence, second ed., Wiley, Hershey, PA, 1992, pp. 330–331.

[76] F. Smarandache, D. Han, A. Martin, Comparative study of contradiction measures in the theory of belief functions, in: Proceedings of the International Conference on Information Fusion (FUSION), IEEE, 2012, pp. 271–277.

[77] S.E. Fahlman, G.E. Hinton, Connectionist architectures for artificial intelligence, Computer 20 (1) (1987) 100–109. http://dx.doi.org/10.1109/MC.1987.1663364.

[78] L. Shastri, A connectionist approach to knowledge representation and limited inference, Cogn. Sci. 12 (3) (1988) 331–392. http://dx.doi.org/10.1207/s15516709cog1203_2.

[79] S. Wrede, J. Fritsch, C. Bauckhage, G. Sagerer, An XML based framework for cognitive vision architectures, in: Proceedings of the International Conference on Pattern Recognition (ICPR), vol. 1, 2004, pp. 757–760. http://dx.doi.org/10.1109/ICPR.2004.1334304.

[80] R. Biswas, S. Thrun, K. Fujimura, Recognizing activities with multiple cues, in: Proceedings of the International Conference on Human Motion: Understanding, Modeling, Capture and Animation, Lecture Notes in Computer Science, vol. 4814, Springer, Berlin, Heidelberg, 2007, pp. 255–270. http://dx.doi.org/10.1007/978-3-540-75703-0_18.

[81] S. Tran, L. Davis, Event modeling and recognition using Markov logic networks, in: D. Forsyth, P. Torr, A. Zisserman (Eds.), Proceeding of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, vol. 5303, Springer, Berlin, Heidelberg, 2008, pp. 610–623. http://dx.doi.org/10.1007/978-3-540-88688-4_45.

[82] M. Tenorth, M. Beetz, KnowRob—Knowledge processing for autonomous personal robots, in: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), IEEE, 2009, pp. 4261–4266. http://dx.doi.org/10.1109/IROS.2009.5354602.

[83] A. Kembhavi, T. Yeh, L. Davis, Why did the person cross the road (there)? Scene understanding using probabilistic logic models and common sense reasoning, in: K. Daniilidis, P. Maragos, N. Paragios (Eds.), Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, vol. 6312, Springer, Berlin, Heidelberg, 2010, pp. 693–706. http://dx.doi.org/10.1007/978-3-642-15552-9_50.

[84] D. Gehrig, P. Krauthausen, L. Rybok, H. Kuehne, U. Hanebeck, T. Schultz, R. Stiefelhagen, Combined intention, activity, and motion recognition for a humanoid household robot, in: Proceedings of the International IEEE Conference on Intelligent Robots and Systems (IROS), IEEE, 2011, pp. 4819–4825. http://dx.doi.org/10.1109/IROS.2011.6095118.

[85] N. Oliver, A. Garg, E. Horvitz, Layered representations for learning and inferring office activity from multiple sensory channels, Comput. Vis. Image Underst. 96 (2) (2004) 163–180. http://dx.doi.org/10.1016/j.cviu.2004.02.004 (Special issue: Event Detection in Video).

[86] L.D. Raedt, Logical and Relational Learning, Cognitive Technologies, vol. XVI, Springer Science & Business Media, 2008.

[87] L. Getoor, B. Taskar, Introduction to Statistical Relational Learning, The MIT press, 2007.

[88] R.P. Mahler, Statistical Multisource–Multitarget Information Fusion, Artech House Inc., Norwood, MA, 2007.

[89] S. Reuter, K. Dietmayer, Pedestrian tracking using random finite sets, in: Proceedings of the International Conference on Information Fusion (FUSION), IEEE, 2011, pp. 1–8.

[90] S. Reuter, K. Dietmayer, S. Handrich, Real-time implementation of a random finite set particle filter, in: H.-U. Heiß, P. Pepper, B.-H. Schlingloff, J. Schneider (Eds.), Sensor Data Fusion: Trends, Solutions, Applications (SDF), Lecture Notes in Informatics, vol. 192, Gesellschaft für Informatik, Berlin, 2011.

[91] S. Reuter, B. Wilking, K. Dietmayer, Methods to model the motion of extended objects in multi-object Bayes filters, in: Proceedings of the International Conference on Information Fusion (FUSION), IEEE, 2012, pp. 527–534.

[92] D. Musicki, R. Evans, Joint integrated probabilistic data association: JIPDA, IEEE Trans. Aerosp. Electron. Syst. 40 (3) (2004) 1093–1099. http://dx.doi.org/10.1109/TAES.2004.1337482.

[93] B. Ristic, B.-N. Vo, D. Clark, B.-T. Vo, A metric for performance evaluation of multi-target tracking algorithms, IEEE Trans. Signal Process. 59 (7) (2011) 3452–3457. http://dx.doi.org/10.1109/TSP.2011.2140111.

[94] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154. http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb.

[95] M. Munz, M. Mahlisch, K. Dietmayer, Generic centralized multi sensor data fusion based on probabilistic sensor and environment models for driver assistance systems, IEEE Intell. Transp. Syst. Mag. 2 (1) (2010) 6–17. http://dx.doi.org/10.1109/MITS.2010.937293.

[96] M. Munz, K. Dietmayer, Using Dempster-Shafer-based modeling of object existence evidence in sensor fusion systems for advanced driver assistance systems, in: IEEE Intelligent Vehicles Symposium (IV), 2011, pp. 776–781. http://dx.doi.org/10.1109/IVS.2011.5940463.

[97] S. Biundo, P. Bercher, T. Geier, F. Müller, B. Schattenberg, Advanced user assistance based on AI planning, Cognit. Syst. Res. 12 (3–4) (2011) 219–236. http://dx.doi.org/10.1016/j.cogsys.2010.12.005 (Special issue: Complex Cognition).

[98] F. Müller, S. Biundo, HTN-style planning in relational POMDPs using first-order FSCs, in: J. Bach, S. Edelkamp (Eds.), Proceedings of the Annual German Conference on Künstliche Intelligenz (KI), Lecture Notes in Computer Science, vol. 7006, Springer, 2011, pp. 216–227. http://dx.doi.org/10.1007/978-3-642-24455-1_20.

[99] D. McDermott, The 1998 AI planning systems competition, AI Mag. 21 (2) (2000) 35–55.

[100] S. Sanner, Relational Dynamic Influence Diagram Language (RDDL): Language Description, Technical Report, NICTA and the Australian National University, 2011.

[101] B. Milch, S. Russell, First-order probabilistic languages: into the unknown, in: S. Muggleton, R. Otero, A. Tamaddoni-Nezhad (Eds.), Proceedings of the International Conference on Inductive Logic Programming (ILP), Lecture Notes in Computer Science, vol. 4455, Springer, Berlin, Heidelberg, 2007, pp. 10–24. http://dx.doi.org/10.1007/978-3-540-73847-3_3.

[102] R. de Salvo Braz, E. Amir, D. Roth, A survey of first-order probabilistic models, in: D. Holmes, L. Jain (Eds.), Innovations in Bayesian Networks Studies in Computational Intelligence, vol. 156, Springer, Berlin, Heidelberg, 2008 http://dx.doi.org/10.1007/978-3-540-85066-3_12.

[103] P. Domingos, D. Lowd, Markov logic: an interface layer for artificial intelligence, Synth. Lect. Artif. Intell. Mach. Learn. 3 (1) (2009) 1–155. http://dx.doi.org/10.2200/S00206ED1V01Y200907AIM007.

[104] T. Geier, S. Reuter, K. Dietmayer, S. Biundo, Goal-based person tracking using a first-order probabilistic model, in: A. Nicholson, J.M. Agosta, M.J. Flores (Eds.), Proceedings of the UAI Bayesian Modeling Applications Workshop (UAI-AW), vol. 962 CEUR-WS.org, CEUR-WS, 2012.

[105] M. Glodek, S. Scherer, F. Schwenker, Conditioned hidden Markov model fusion for multimodal classification, in: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), ISCA, 2011, pp. 2269–2272.

[106] M. Glodek, F. Schwenker, G. Palm, Detecting actions by integrating sequential symbolic and sub-symbolic information in human activity recognition, in: P. Perner (Ed.), Proceedings of the International Conference on Machine Learning and Data Mining (MLDM), Lecture Notes in Computer Science, vol. 7376, Springer, Berlin, Heidelberg, 2012, pp. 394–404. http://dx.doi.org/10.1007/978-3-642-31537-4_31.

[107] S. Ultes, R. ElChabb, A. Schmitt, W. Minker, JaCHMM: A Java-based conditioned hidden Markov model library, in: Proceeding of the International IEEE conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 3213–3217.

[108] M. Glodek, T. Geier, S. Biundo, F. Schwenker, G. Palm, Recognizing user preferences based on layered activity recognition and first-order logic, in: Proceedings of the International IEEE Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2013, pp. 648–653.

[109] R. Sharma, V. Pavlovic, T. Huang, Toward multimodal human–computer interface, Proc. IEEE 86 (5) (1998) 853–869. http://dx.doi.org/10.1109/5.664275.

[110] J.R. Carbonell, Mixed-initiative man-computer instructional dialogues (Ph.D. thesis), Department of Electrical Engineering of the Massachusetts Institute of Technology, URL ⟨http://dspace.mit.edu/handle/1721.1/13801⟩, 1970 (last visited 01/10/2013).

[111] R.A. Bolt, "Put-that-there": voice and gesture at the graphics interface, Comput. Graph. 14 (3) (1980) 262–270.

[112] J.G. Neal, S.C. Shapiro, Intelligent multi-media interface technology, ACM SIGCHI Bull. 20 (1) (1988) 11–41. http://dx.doi.org/10.1145/49103.1046407.

[113] W. Wahlster, User and discourse models for multimodal communication, in: J.W. Sullivan, S.W. Tyler, J.W. Sullivan, S.W. Tyler (Eds.), Readings in Intelligent User Interfaces, ACM Press, 1991, pp. 45–67. http://dx.doi.org/10.1145/107215.128691.

[114] S. Bangalore, M. Johnston, Integrating multimodal language processing with speech recognition, in: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), vol. 2, 2000, pp. 126–129.

[115] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, S. Feiner, Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality, in: Proceedings of the International Conference on Multimodal Interfaces (ICMI), ACM, New York, NY, 2003, pp. 12–19. http://dx.doi.org/10.1145/958432.958438.

[116] S. Oviatt, Mutual disambiguation of recognition errors in a multimodel architecture, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, 1999, pp. 576–583. http://dx.doi.org/10.1145/302979.303163.

[117] S. Larsson, D.R. Traum, Information state and dialogue management in the TRINDI dialogue move engine toolkit, Nat. Lang. Eng. 6 (3–4) (2000) 323–340. http://dx.doi.org/10.1017/S1351324900002539.

[118] N. Nguyen, D. Phung, S. Venkatesh, H. Bui, Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, IEEE, 2005, pp. 955–960. http://dx.doi.org/10.1109/CVPR.2005.203.

[119] J.D. Williams, S. Young, Partially observable Markov decision processes for spoken dialog systems, Comput. Speech Lang. 21 (2) (2007) 393–422. http://dx.doi.org/10.1016/j.csl.2006.06.008.

[120] C. Lee, S. Jung, K. Kim, D. Lee, G.G. Lee, Recent approaches to dialog management for spoken dialog systems, J. Comput. Sci. Eng. 4 (1) (2010) 1–22.

[121] S. Young, J. Schatzmann, K. Weilhammer, H. Ye, The hidden information state approach to dialog management, in: Proceedings of the International IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, IEEE, 2007, pp. 149–152. http://dx.doi.org/10.1109/ICASSP.2007.367185.

[122] A. Nguyen, W. Wobcke, An agent-based approach to dialogue management in personal assistants, in: Proceedings of the International Conference on Intelligent User Interfaces (IUI), ACM, New York, NY, 2005, pp. 137–144. http://dx.doi.org/10.1145/1040830.1040865.

[123] D. Costa, C. Duarte, Adapting multimodal fission to user's abilities, in: Proceedings of the International Conference on Universal Access in Human-Computer Interaction (UAHCI): Design for all and eInclusion—Part I, Lecture Notes in Computer Science, vol. 6765, Springer, Berlin, Heidelberg, 2011, pp. 347–356. http://dx.doi.org/10.1007/978-3-642-21672-5_38.

[124] M.E. Foster, State of the Art Review: Multimodal Fission, Public Deliverable 6.1, University of Edinburgh, URL ⟨http://groups.inf.ed.ac.uk/comic/documents/deliverables/Del6-1.pdf⟩, 2002 (last visited 01/10/2013).

[125] C. Rousseau, Y. Bellik, F. Vernier, D. Bazalgette, A framework for the intelligent multimodal presentation of information, Signal Process. 86 (12) (2006) 3696–3713. http://dx.doi.org/10.1016/j.sigpro.2006.02.041.

[126] B. Dumas, D. Lalanne, S. Oviatt, Multimodal interfaces: A survey of principles, models and frameworks, in: D. Lalanne, J. Kohlas (Eds.), Human Machine Interaction—Research Results of the MMI Program, Lecture Notes in Computer Science, vol. 5440, Springer, Berlin, Heidelberg, 2009, pp. 3–26. http://dx.doi.org/10.1007/978-3-642-00437-7_1.

[127] M.D. Hina, C. Tadj, A. Ramdane-Cherif, N. Levy, A multi-agent based multimodal system adaptive to the user's interaction context, in: Multiagent Systems, InTech, 2011, pp. 29–56 (Chapter 2). http://dx.doi.org/10.5772/14692.

[128] D. Roscher, M. Blumendorf, S. Albayrak, A meta user interface to control multimodal interaction in smart environments, in: Proceedings of the International Conference on Intelligent User Interfaces (IUI), ACM, New York, NY, 2009, pp. 481–482. http://dx.doi.org/10.1145/1502650.1502725.

[129] P. Smets, Data fusion in the transferable belief model, in: Proceedings of the International Conference on Information Fusion (FUSION), vol. 1, IEEE, 2000, pp. PS21–PS33. http://dx.doi.org/10.1109/IFIC.2000.862713.

[130] F. Schüssel, F. Honold, M. Weber, Using the transferable belief model for multimodal input fusion in companion systems, in: F. Schwenker, S. Scherer, L.-P. Morency (Eds.), Proceeding of the ICPR 2012 Satellite Workshop on Multimodal Pattern Recognition of Social Signals in Human Computer Interaction (MPRSS), Lecture Notes in Computer Science, vol. 7742, Springer, Berlin, Heidelberg, 2013, pp. 100–115. http://dx.doi.org/10.1007/978-3-642-37081-6_12.

[131] U. Brandes, M. Eiglsperger, I. Herman, M. Himsolt, M. Marshall, GraphML progress report: structural layer proposal, in: P. Mutzel, M. Jünger, S. Leipert (Eds.), Proceedings of the International Symposium on Graph Drawing (GD), Lecture Notes in Computer Science, vol. 2265, Springer, Berlin, Heidelberg, 2002, pp. 501–512. http://dx.doi.org/10.1007/3-540-45848-4_59.

[132] F. Nothdurft, G. Bertrand, T. Heinroth, W. Minker, GEEDI—Guards for emotional and explanatory dialogues, in: Proceedings of the International Conference on Intelligent Environments (IE), 2010, pp. 90–95. http://dx.doi.org/10.1109/IE.2010.24.

[133] F. Nothdurft, G. Bertrand, H. Lang, W. Minker, Adaptive explanation architecture for maintaining human–computer trust, in: Proceedings of the IEEE Computer Software and Applications Conference (COMPSAC), 2012, pp. 176–184. http://dx.doi.org/10.1109/COMPSAC.2012.28.

[134] R. Barták, Constraint programming: In pursuit of the holy grail, in: Proceedings of the Week of Doctoral Students (WDS), vol. IV, MatFyzPress, 1999, pp. 555–564.

[135] A.J. Fernández, T. Hortalá-González, F. Sáenz-Pérez, R. Del Vado-Vírseda, Constraint functional logic programming over finite domains, Theory Practice Logic Program. 7 (5) (2007) 537–582. http://dx.doi.org/10.1017/S1471068406002924.

[136] G. Bertrand, F. Nothdurft, W. Minker, "What do you want to do next?" providing the user with more freedom in adaptive spoken dialogue systems, in: Proceedings of the International Conference on Intelligent Environments (IE), 2012, pp. 290–296. http://dx.doi.org/10.1109/IE.2012.27.

[137] F. Honold, F. Schüssel, M. Weber, F. Nothdurft, G. Bertrand, W. Minker, Context models for adaptive dialogs and multimodal interaction, in: Proceedings of the International Conference on Intelligent Environments (IE), IEEE, 2013. http://dx.doi.org/10.1109/IE.2013.54.

[138] F. Honold, F. Schüssel, M. Weber, The automated interplay of multimodal fission and fusion in adaptive HCI, in: 2014 10th International Conference on Intelligent Environments (IE), IEEE, Shanghai, China, 2014, pp. 170–177. http://dx.doi.org/10.1109/IE.2014.32.

[139] P. Bercher, S. Biundo, T. Geier, T. Hoernle, F. Nothdurft, F. Richter, B. Schattenberg, Plan, repair, execute, explain—how planning helps to assemble your home theater, in: Proceedings of the 24th International Conference on Automated Planning and Scheduling (ICAPS), AAAI Press, 2014, pp. 386–394.

[140] F. Schüssel, F. Honold, M. Weber, Influencing factors on multimodal interaction during selection tasks, J. Multimodal User Interfaces 7 (4) (2013) 299–310. http://dx.doi.org/10.1007/s12193-012-0117-5.

[141] C. Britton, P. Bye, IT Architectures and Middleware: Strategies for Building Large, Integrated Systems, Addison-Wesley Professional, 2004.

[142] J.M. Myerson, The Complete Book of Middleware, Auerbach Publications, 2002.

[143] A. Puder, K. Römer, F. Pilhofer, Distributed Systems Architecture: A Middleware Approach, The MK/OMG Press Series, Morgan Kaufmann, 2006.

[144] M. Schröder, The SEMAINE API: towards a standards-based framework for building emotion-oriented systems, Adv. Hum. Comput. Interact. 2010 (2010) 1–21. http://dx.doi.org/10.1155/2010/319406.

Michael Glodek has graduated in Computer Science from the University of Lübeck, Germany and is working since 2009 as a Research Assistant at the University of Ulm, Germany as part of the Transregional Collaborative Research Centre SFB/TRR 62 Companion-Technology for Cognitive Technical Systems. His research covers artificial intelligence, neural information processing and machine learning. His recent studies focus on probabilistic models for the integration of sub-symbolic and symbolic information in cognitive technical systems.

Günther Palm was born in 1949. He studied Mathematics in Hamburg and Tübingen. After his graduation in Mathematics (Master in 1974, Ph.D. in 1975) he worked at the MPI for Biological Cybernetics in Tübingen on the topics of nonlinear systems, associative memory and brain theory. In 1983/1984 he was a Fellow at the Wissenschaftskolleg in Berlin. From 1988 to 1991 he was a Professor for theoretical brain research at the University of Düsseldorf. Since then he is a Professor for Computer Science and the Director of the Institute of Neural Information Processing at the University of Ulm. He is working on information theory, associative memory, pattern recognition, neural networks, and brain modelling.

Friedhelm Schwenker received his Ph.D. degree in Mathematics from the University of Osnabrück in 1988. From 1989 to 1992 he was a Postdoc at the Vogt-Institute for Brain Research at the Heinrich Heine University Düsseldorf. Since 1992 he is a Researcher and Senior Lecturer at the Institute of Neural Information Processing at the University of Ulm. His main research interests are artificial neural networks, machine learning, data mining, pattern recognition, applied statistics and affective computing.

Frank Honold has a degree in Computer Science and graduated from the Institute of Media informatics, Ulm University. Since 2009 he is a Research Assistant and Member of the scientific staff. He is doing his Ph.D. in the Transregional Collaborative Research Center SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” which is funded by the German Research Foundation (DFG). His thematic focus is on the topics of multimodal fission and modality arbitration in the domain of multimodal systems.

Felix Schüssel studied Media Informatics at the University of Ulm. After graduation, he worked for the Fraunhofer Institute for Experimental Software Engineering (IESE) at Daimler AG, Ulm. Since 2009 he is a Ph.D. student and Research Assistant at the Institute for Media Informatics at the University of Ulm. Topics include the fusion of user inputs with uncertainty and the abstract modeling of interactions.

Michael Weber holds a Ph.D. in Computer Science from the University of Kaiserslautern. After some years in industry he joined the University of Ulm as a Professor for Computer Science in 1994 and was appointed as the Director of the Institute of Media Informatics in 2000. His current research interests include mobile and ubiquitous computing systems and human-computer interaction.

Thomas Geier is a Ph.D. student at the Institute of Artificial Intelligence, Ulm University since 2009. He is working on approximate inference methods for probabilistic, graphical models. Previously, he was working as a Software Developer for embedded systems, after obtaining his Diploma in Computer Science from the University of Würzburg, Germany in 2007.

Thilo Hörnle studied Computer Science at the University of Ulm. Since 2009 he works in the project “Data management and system integration” of the SFB/Transregio 62 at the University of Ulm. His research interest is the development of a general abstract architecture for Companion systems.

Susanne Biundois a Professor of Computer Science at Ulm University. She is the chair of the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems?. In the Centre, more than 70 scientists from the areas of informatics, electrical engineering, medicine, neurobiology, and psychology systematically explore cognitive abilities and their implementation within technical systems. Susanne Biundo was the initiator and coordinator of PLANET, the "European Network of Excellence in AI Planning?. Her research interests include AI Planning, Automated Reasoning, Knowledge Modeling, and Cognitive Systems.

Gerald Krell is a Lecturer at the Institute for Information und Communication Technology (IIKT) of Otto-von-Guericke University Magdeburg, Germany. Main topics of research span the use of neural networks in image processing, image coding and data fusion with applications in medicine, man–machine interaction and industry.

Florian Nothdurft received his diploma degree in Computer Science from the Ulm University, Germany, in 2010. He is currently pursuing the Ph.D. degree as a member of the Dialogue Systems Group at the Institute of Communications Engineering in Ulm. His current research focus lies on how to impart user- and situation adaptive explanations in dialogue systems.

Wolfgang Minker is a Professor at the University of Ulm, Department of Communications Technology, Germany. He received his Ph.D. in Engineering Science from the Karlsruhe Institute of Technology (formerly University of Karlsruhe), Germany, in 1997 and his Ph.D. in Computer Science from the University of Paris-Sud, France, in 1998. He was a Researcher at the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'ingénieur (LIMSI-CNRS), France, from 1993 to 1999 and a member of the scientific staff at DaimlerChrysler, Research and Technology, Germany, from 2000 to 2002.

Stephan Reuter was born 1981 in Crailsheim, Germany. He received a Diploma degree (equivalent to M.Sc. degree) and the Dr.-Ing, degree (equivalent to Ph.D.) in Electrical Engineering from Ulm University in 2008 and 2014, respectively. Currently, he has a Post Doc position at the Institute of Measurement, Control and Microtechnology in the school of Engineering and Computer Science at Ulm University. His main research topics are sensor data fusion, multi-object tracking, environment perception for cognitive technical systems, and sensor data processing.

Klaus Dietmayer (M'05) was born in Celle, Germany in 1962. He received his Diploma degree (equivalent to M.Sc. degree) in 1989 in Electrical Engineering from the Technical University of Braunschweig (Germany), and the Dr.-Ing. degree (equivalent to PhD) in 1994 from the University of Armed Forces in Hamburg (Germany). In 1994 he joined the Philips Semiconductors Systems Laboratory in Hamburg, Germany as a Research Engineer. Since 1996 he became a manager in the field of networks and sensors for automotive applications. In 2000 he was appointed to a Professorship at Ulm University in the field of measurement and control. Currently he is a Full Professor and the Director of the Institute of Measurement, Control and Microtechnology in the school of Engineering and Computer Science at Ulm University. His research interests include information fusion, multi-object tracking, environment perception for advanced automotive driver assistance, and E-Mobility.