Fusion of perceptual processes for real-time object tracking

Kai Jüngling, Michael Arens
Research Institute for Optronics and Pattern Recognition
FGAN-FOM
76275 Ettlingen, Germany
Email: {juengling,arens}@fom.fgan.de

Marc Hanheide, Gerhard Sagerer
Applied Computer Science
Bielefeld University
33501 Bielefeld, Germany
Email: {mhanheid,sagerer}@techfak.uni-bielefeld.de

***Abstract*— This paper introduces a generic architecture for the fusion of perceptual processes and its application in real-time object tracking. In this architecture, the well known anchoring approach is, by integrating techniques from information fusion, extended to multi-modal anchoring so as to be applicable in a multi-process environment. The system architecture is designed to be applicable in a generic way, independent of specific application domains and of the characteristics of the underlying sensory processes. It is shown that, by combining multiple independent video-based detection methods, the generic multi-modal anchoring approach can be successfully employed for real-time person tracking in difficult environments.**

Keywords: Object tracking, person tracking, anchoring, multi-modal anchoring, fusion architecture

I. Introduction

Fusion of information from different sources is a technique that we use in everyday life without even being aware of it. Humans use different (multi-modal) senses simultaneously to generate a detailed image of the environment by combining perceptual information from different "sensory" sources.

Using multiple sensory sources has a lot of advantages compared to using just one information source. As detailed in [6] and [16], these advantages comprise higher robustness, larger spatial and temporal coverage, increased overall reliability, reduced ambiguity and availability of complementary information. Considering these advantages, it is highly desirable to make use of multiple information sources whenever possible. While fusing information seems to be very easy to accomplish for humans, the transfer of this ability to an artificial system involves considerable difficulty.

Since the information is provided by a sensor, the system has to deal with *uncertain information*. We thus need to model the uncertainty in the single processes and find a way to combine these uncertainties to generate a joint hypothesis. Sensors that provide different kinds of information are likely to work at different frequencies. When fusing information from these sensors, we have to find a way to deal with *incomplete information*. Depending on the sensor properties and the kind of information the sensor provides, even a single sensor need not necessarily provide data at a constant rate. This implies that the system has to deal with *asynchronous information*, even inside a single process. When looking at sensory processes in a multi-modal environment, it becomes clear that these processes provide different kinds of information, possibly with information content and depth even on different levels. These kinds of information are not inherently compatible, which leads to the problem of *combining inhomogeneous data*.

The objective of this work is to provide a generic architecture that can be used to combine multiple perceptual processes under these circumstances. The development of the architecture focuses mainly on applicability in real-time object tracking. For this purpose, Sec. II looks at the anchoring framework, which describes a general object tracking approach in a single sensor domain. Furthermore, the extensions necessary in a multi-sensor environment are introduced on a theoretical basis, which leads to the notion of multi-modal anchoring. Sec. III takes a deeper look at the application environment, especially at the characteristics of the perceptual processes. Sec. IV describes the system architecture that implements the anchoring framework and integrates the components needed due to the extensions described in II. Sec. V describes the application of the system for person tracking and presents the results.

II. Anchoring symbols to sensor data

The anchoring framework, introduced by S. Coradeschi and A. Saffiotti in [1] and [2], provides a method for tracking objects over time by defining the theoretical basis for grounding symbols to sensor data. Formally, they describe anchoring as "the process of creating and maintaining the correspondence between symbols and percepts that refer to the same physical object". This correspondence is represented by the so called anchor, which establishes the connection between the symbolic and sensory levels. Formally, the anchor is any partial function α from time to triples in $\chi \times \Pi \times (\Phi \rightarrow D(\Phi))$, where

- $\chi = \{x_1, x_{2,}, ...\}$ is a set of *symbols*.
- $\Pi = \{\pi_1, \pi_2, ...\}$ is a set of *percepts*. A percept is a structured collection of measurements assumed to originate from the same physical object. The measurements are described by a set of attributes $\Phi = \{\phi_1, \phi_2, ...\}$.
- The anchor *signature* $\gamma : \Phi \rightarrow D(\Phi)$ is a partial function of the attribute set Φ to values of the domain $D(\Phi) = \bigcup_{\phi \in \Phi} D(\phi)$.

An important component of the anchoring framework is the symbolic description of objects by predicates. Starting from this description, anchors that fulfill the *grounded* definition are searched in incoming data in a top-down manner. An anchor is *grounded* at time t, if the percept assigned to this anchor α is perceived at time t and is also valid for the symbolic description of α.

Since the original anchoring approach deals only with a single percept type and precisely one perceptual process, it has to be extended to fit our needs in a multi-process environment. This leads to the idea of multi-modal anchoring.

Previous work on multi-modal anchoring was done by Kleinhagebrock et. al. in [4] and Lang et. al. in [5], [3]. They split up the anchoring process into individual component anchoring processes, one for each perceptual process. The *composite anchor* then constitutes a common description by merging the component anchors. In contrast to their approach, we do not split up the anchoring process into subprocesses, but extend the anchoring approach itself to merge multiple perceptual processes in a single anchoring process. Following an idea of Hanheide [7], we generalize the anchoring approach by integrating concepts known from information fusion. In additional, a reliability value that serves to model the uncertainty in both, the percepts and the anchor hypotheses is included into the anchoring process. Specifically, our approach extends the anchor definition to link multiple percepts to a single symbol. The anchor is defined as a triple, consisting of a symbol $x \in \chi$, a set of percepts $\Lambda \subseteq \Pi$, and a signature Γ that is calculated by combining the attributes that are comprised in the set of percepts. Here,

- $\chi = \{x_{1,}x_{2,},...\}$ denotes the symbol set again, while
- $\Pi = \bigcup_{i \in I} \Pi_i$ is the union of the percepts of all perceptual processes $\Pi_i = \{\pi_{i,1}, \pi_{i,2}, ...\}$ and
- $\Gamma \subseteq \bigcup_{i \in I} \Phi_i$ stands for the anchor signature that includes a subset of the attribute set, whereby Φ_i constitutes the attributes of the i-th percept.

Additionally, the anchor signature is extended by the reliability value $\rho \in [0, 1]$. This value reflects the amount of confidence we give to the anchor hypothesis, whereby 0 is the minimum, 1 is the maximum possible confidence. The percept definition is extended as well. The extended percept $\pi \in \Pi$ comprises a set of attributes Φ^{per} that describe the properties of the referenced object; a timestamp t^{per} that determines the time of perception; and the reliability value ρ^{per}, that is assigned by the perceptual process from which the percept originates.

Since multi-modal anchoring deals with multiple percepts for a single anchor hypothesis that are themselves treated as hypotheses with a certain reliability, the original grounded definition does not suffice for our needs. The *grounded* definition needs to be adapted by including the newly introduced anchor reliability ρ^{anc}. An anchor is *grounded* at a time t, when its reliability ρ^{anc} exceeds a certain *grounded threshold* $\rho^{grounded}$. An example how this threshold is determined on training data in a specific application is given in Subsec. V-C.

Since the original predicate grounding relation is still applicable to establish the connection between the signature's attribute values and a symbolic description by predicates, multi-modal anchoring can serve as the basis for symbolic higher level reasoning processes just as well as the original approach does. Since reliability information is available for the anchor hypothesis as well as the anchors attributes, it can directly be used to represent uncertainty on the symbolic level.

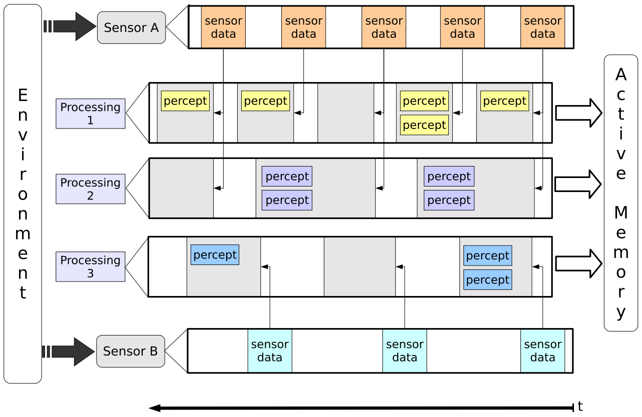

Figure 1. Perceptual processes

III. System Environment: Perceptual processes

Our object tracking system works in a multi-sensor, or more abstractly, a multi-process environment which includes the difficulties outlined in Sec. I.

Figure 1 shows a sample system environment. In this case, two sensors acquire environmental data at different rates. The data provided by ”Sensor A” triggers two different signal processing methods that provide a set of percepts for each sensor data set they process. Since the processing time of the sensor data differs depending on the complexity of the underlying algorithms, the working rates of the processing methods can differ too. A single processing method is applied to the data of ”Sensor B”. Although this method works at the full sensor rate, it is not synchronous with ”Processing 1” because the sensor rates differ. This means that the anchoring system has to deal with asynchronous and asymmetrical incoming percepts. Since the main components of the anchoring system work on an abstract level, it is irrelevant whether these percepts originate from different sensors or just from different processing methods applied to the data of a single sensor. Hence, from now on, we use the abstract term *perceptual process* to refer to any kind of external process that provides the system with percepts.

These processes must fulfill two basic properties to be compatible with our multi-modal anchoring approach. It is assumed that, at every processing step of a perceptual process, the result of this process is a *set of percepts*, which can be distinguished by their attribute values. The level of detail of the information provided by different percept types can however vary significantly. For instance, the system can integrate results of an object detection process that provides high-level information about object characteristics and class membership as well as results of a simple motion detection method that provides coordinates where motion has been detected. Furthermore, each percept has a *reliability* value assigned by the perceptual process that can be used as a confidence measure in the fusion process.

Although we regard all processes that fulfill these requirements as perceptual processes with which our system can interface, we distinguish two kinds of processes based on another criterion.

Data driven processes are triggered directly by sensor data and establish a one-way communication with the anchoring system by providing it with percepts. Since our system should provide a way to incorporate system runtime knowledge into perceptual processes, we introduce another kind of process that establishes bidirectional communication between the perceptual level and the anchoring system.

These *expectation driven* processes are started at runtime by the anchoring system in a top-down manner for a specific system entity. Since the initialization should be an automatic process, we must provide a way to define when this should happen. This is done by involving predefined model assumptions and conditions. These can employ things like the reliability of an anchor hypothesis or a specific attribute, the hypothesis' age, or model assumptions that define constraints for attribute values. Expectation driven percepts contain an additional "ID" field that identifies the entity inside the anchoring system the process has been started for. The main difference in the processing of these two kind of processes inside the anchoring system occurs at the assignment step, that is described in Subsec. IV-A. For expectation driven processes, no matching step is needed to determine the assignment to an anchor hypothesis, because this assignment is already known. However, the match can be used to ascertain the correct functionality of the underlying process and, if necessary, terminate or reinitialize this process.

In principle, a vast number of different implementations of expectation driven processes are imaginable. These can be processes that are initialized by the system with runtime knowledge and then provide percepts in the same way data-driven processes do, as well as processes that provide specific information for a single portion of sensor data, e.g., to verify anchor hypotheses or acquire detailed information. An additional possible application is the use of expectation driven processes to directly control the sensory level (for example pan or zoom a camera to an area of interest).

The flow of information in the communication of perceptual processes and the anchoring system is inherently asynchronous and parallel. Following ideas derived from human memory organization [17], the active memory concept [8] is applied to integrate the different processes and fusion in a generic XML-based integration architecture [9]. The underlying concept of chunking information XML documents has been published in previous work [10].

IV. System Architecture

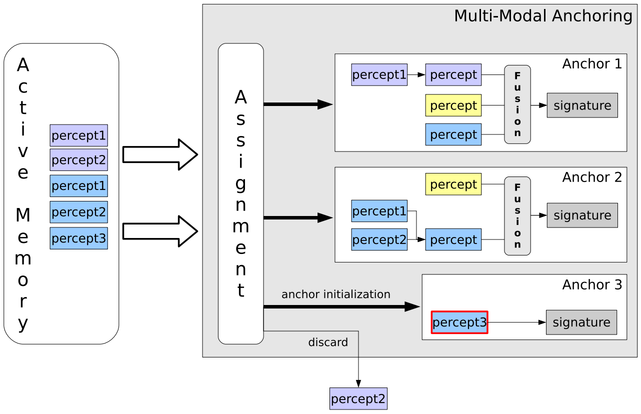

Figure 2. Overview of the process flow in the multi-modal anchoring system.

As shown in Fig. 2, the multi-modal anchoring consists of two steps that are triggered by incoming percepts. The first step is the *assignment of percepts*. (Note that the percepts need not necessarily enter the anchoring system at the same time.) Three different options are considered in this process. A percept is either assigned to an existing anchor hypothesis, discarded or used to initialize a new anchor hypothesis. Which option applies depends on the result of a matching between existing anchor hypotheses and the percept, as well as the reliability value of the percept. The details of the assignment are described below in Subsec. IV-A. In case an incoming percept has generated a new anchor hypothesis or has been assigned to an existing one, the new information provided by the percept must be incorporated into the hypothesis. For this purpose, the signature of the anchor hypothesis is recalculated in a two-step update process. This is detailed in Subsec. IV-B.

A. *Percept Assignment*

The most important step inside an object tracking system is to determine which part of the data acquired by a sensor belongs to which entity inside the system. In anchoring, this is accomplished by calculating an assignment between incoming percepts and existing anchor hypotheses.

1) Matching: The first step of the assignment is to determine the likelihood that a certain percept was generated by the object to which a specific anchor hypothesis refers. This is done by a matching procedure. As described in Sec. II, the percept is composed of a set of attributes Φ. Starting from these attributes, the matching procedure calculates the similarity of an anchor hypothesis and a percept as follows.

For each attribute $\phi_i \in \Phi$, one checks whether the anchor signature contains an attribute of a compatible type. This is done by a function Ψ that assigns a compatible signature attribute $\gamma_k \in \Gamma$ to each percept attribute or assigns 0, if no compatible attribute is available:

$$\Psi : (\Phi \times \Gamma) \rightarrow \{\Phi \cup 0\}. \tag{1}$$

Compatible in this case refers to compatibility integrated by a transfer function κ, which can perform any transformation between a $\phi_i \in \Phi$ and a $\gamma_k \in \Gamma$. This transformation could be a simple offset integration like we use to incorporate a face percept in a person model or a function that converts between different coordinate systems as used in [3]. If percept and anchor do not contain any common attributes, Γ is 0 for all $\phi_i \in \Phi$ and no similarity can be defined. Otherwise, the overall match of an anchor hypothesis and a percept is defined by:

$$\beta(\Phi, \Gamma, \Psi) = \frac{\sum_{i=0}^{I} \omega_i \beta_i(\kappa(\phi_i), \Psi(\phi_i, \Gamma))}{\sum_{i=0}^{I} \omega_i}. \tag{2}$$

This matching involves comparison of the individual attributes on the basis of the assignment Ψ. This is done by the type-specific functions β_i, the implementation of which depends on the attribute domains and is part of the application-specific components of the multi-modal anchoring. The weights ω_i reflect the importance of the single attributes in the matching process. By these weights, attributes that are more significant for the calculation of similarity can be weighted more than attributes that are not applicable as similarity measures. Attributes that shouldn't be used in the matching process can be defined as *passive* by setting the weight to 0. A special case of an attribute is the *exclusion attribute*. If the match value of an exclusion attribute is below that attribute's internal threshold, assignment of the percept to the according anchor is prohibited. This exclusion strategy makes sense for attributes that can disqualify the compatibility of percept and anchor. For example a percept attribute that defines object class "face", can disqualify the assignment to an anchor that references a car object.

If $\beta(\Phi, \Gamma, \Psi) = 1$, percept and anchor hypotheses are considered a perfect match; if $\beta(\Phi, \Gamma, \Psi) = 0$, they do not match at all.

2) Assignment models: The match measurement β provides a general criterion by which the anchoring process can decide (i) which percepts are assigned to which anchor hypotheses, (ii) which percepts are used to initialize a new hypothesis, and (iii) which percepts are discarded completely.

The principle assignment task is defined as:
Determine the best assignment of a set of I percepts $\Pi = \{\pi_1, .., \pi_I\}$ to a set of K anchor hypotheses $\Lambda = \{\alpha_1, ..., \alpha_{K-2}, \mu, \nu\}$ based on a similarity measure $\delta_{i,k} = \delta(\pi_i, \alpha_k)$. Here, the percepts Π result from one perceptual process and have been perceived at the same time. The virtual anchors μ and ν represent anchor initialization and discarding of a percept, respectively. The assignment to these anchors is decided on the basis of thresholds applied on match measure β and percept reliability ρ. Since the assignment requirements (and thus the definition of the best assignment) vary for different percept types, two assignment strategies are integrated into the anchoring system. The first strategy is a very simple one that allows for independent processing of each $\pi_i \in \Pi$. The percepts are processed sequentially and each percept is assigned to the anchor hypothesis with the highest match value. The main advantages of this approach are low time complexity and the ability to independently process the percepts that were generated at the same time by the same process. This introduces the possibility of intra-process parallelization. A disadvantage is that it is not possible to exclusively assign percepts and calculate the best assignment at the same time when processing the percepts sequentially. Hence, this approach is useful only when it is allowed to *assign multiple percepts to a single anchor hypothesis (N:1)*. Mutual exclusion is desired, e.g., in the case where percepts represent faces and anchors reference persons. In this case, assignment of multiple percepts to the same anchor should be avoided for obvious reasons. When this is demanded, a more sophisticated assignment strategy must be applied to ensure the best *assignment* in an *exclusive (1:1)* way.

In this case, the best assignment is defined as the assignment of I percepts to K anchor hypotheses

$$\begin{aligned} & i \rightarrow \eta(i), 1 \leq i \leq I, 1 \leq \eta(i) \leq K \\ & i \neq j \Rightarrow \eta(i) \neq \eta(j) \end{aligned} \tag{3}$$

that maximizes the sum of similarities:

$$\delta_{tot} = \sum_{i=0}^{I} \delta(i, \eta(i)) \tag{4}$$

This task can be regarded as the problem of searching for the maximum weight in a weighted bipartite graph, which is known as the assignment problem. A well known approach to solving this problem in polynomial time is the Hungarian algorithm presented in [12] by H. Kuhn and revised by J. Munkres in [11].

Regardless of the assignment strategy used, assignments that are not supported by a match value beyond a heuristically determined threshold are discarded. The percepts involved in these assignments are formally assigned to the virtual anchors μ or ν.

B. Hypothesis update

After a new percept has been assigned to an existing anchor hypothesis, the new information should be integrated into the hypothesis. This is done by fusing the information available in the anchor signature with the new information carried by the percept.

The fusion is realized in a two-step procedure that is designed to incorporate new information as soon as it is available, without synthetically align or synchronize the perceptual processes inside the anchoring system. This approach prevents the system from being slowed down by the process with the lowest working frequency. By subdividing the fusion process into two parts, the system is also inherently stable against temporal distortions between perceptual processes, which is a main concern in a system that uses distributed processes.

1) Temporal fusion: The temporal fusion treats every perceptual channel (different channels inside the anchoring system refer to different perceptual processes) separately and performs the intra-channel fusion of percepts over time. The motivation for this approach is to gain higher stability in even a single channel and therefore to be more robust to noise and failures inside the perceptual processes. Since the multi-modal anchoring approach is desired to be applicable in a generic way, we must account for different process characteristics. Thus, to gain the best possible adaptivity of the fusion process to percept specific requirements, different fusion strategies are possible. Since these strategies depend on the specific attribute types, the attribute fusion is decoupled from the base system and part of the application-specific component. The fusion of the reliability value (evidence combination) is identical for all attributes and can thus be described generically for a new percept and an old percept that is already included in the anchor.

When combining these percepts, it is important to include the time information of both, the old and the new percept. This is done for two reasons. First, it is possible that more than one percept of the same process and time is assigned to an anchor. These percepts should be treated in a different way than percepts that were perceived at different points in time. Second, the age of the old percept should be incorporated into the fusion process, because it is a measure for the relevance of this information at the present point in time.

Considering these things, the new reliability $\hat{\rho_T}$ is calculated as follows:

$$\hat{\rho_T} = \frac{\sum_{i=0}^{I} \rho_{T,i} + \omega \rho_{old}}{1+\omega}. \tag{5}$$

The reliability at time t is thus defined as the sum of the reliabilities of the percepts, perceived at that time $\rho_{t,i}$ and the weighted reliability of the old percept ρ_{old}. Here, the weight determines the influence of the old percept and is defined by

$$\omega = \beta\epsilon(t_{\rho_T} - t_{\rho_{old}}), \tag{6}$$

where β is the match of percept and anchor hypothesis. This factor is involved in order to let percepts with high match value benefit from high reliability of previous percepts. Using this approach, short time failures inside the perceptual processes can be compensated. Since strong new hypotheses should be inserted with their full expressiveness, the weight is set to 0 if the reliability of the new percept exceeds the reliability of the old percept.

$\epsilon(t_{\rho_T} - t_{\rho_{old}})$ integrates the time weight of the old percept by

$$e(t) = \begin{cases} e^{1-(\frac{t}{t_{exp}})} & \text{if } t > t_{exp} \\ 1 & \text{else} \end{cases} \tag{7}$$

Here, t_{exp} is a process specific constant that defines the expected update interval of the perceptual process. If a hypothesis gains support in less than the time given by t_{exp}, $\epsilon(t)$ evolves to 1. Otherwise the weight decreases because the relevance of the old information also decreases with ongoing time.

2) Process fusion: After the new information has been integrated into the anchor hypothesis by temporal fusion, the information must be integrated into the anchor signature Γ. This is done by fusing the information of all perceptual channels available for this hypothesis. For every signature attribute γ_k, all compatible attributes are combined to calculate the best estimate of the attribute value. Since the fusion is attribute-type-independent and follows a general scheme, it can be described generically for all attribute types by

$$\gamma_k = \frac{\sum_{i=0}^{I} \omega_i \kappa(\phi_i)}{\sum_{i=0}^{I} \omega_i}, \tag{8}$$

where ϕ_i is the i-th attribute of the set of compatible attributes for γ_k.

The weighting factor

$$\omega_i = \epsilon_i \rho_i \tau_i \zeta_i \tag{9}$$

specifies the influence of the single percept types and is composed of the factors

- $\epsilon_i = \epsilon(t^{up} - t_i^{gen})$: Determines the influence of the attribute's age. This factor is 1 for new percepts and 0 for percepts, the perception of which took place so long ago that the information they provide is not useful any more. This factor depends on the current time t^{up}, the time t^{gen} when the percept to which attribute ϕ_i belongs was perceived, and the already introduced constant t_{exp}.
- ρ_i : The reliability of the percept to which attribute ϕ_i belongs.
- τ_i : The weight of the perceptual process in which ϕ_i was perceived. This weight is determined in an a priori training step based on an annotated training set. In addition, it is adapted at runtime using a system-internal process-quality measure that is described in Subsec. IV-C.
- ζ_i : An attribute-specific weighting factor that allows the integration of model knowledge. A simple example for this integration in the context of person tracking is to set the weight of detected foreground regions dependent on the spatial similarity to a person model.

3) Reliability update: In contrast to the signature's attribute values, the signature's reliability (on attribute and anchor level) is not only updated when new information is present but also decreases steadily when no new support for a hypothesis is available.

This is done by

$$\rho_A = \sum_{i=0}^{I} \rho_i \epsilon_i \tau_i. \tag{10}$$

In case of anchor reliability ($\rho_A = \rho_\alpha$), the sum is formed over the *I* perceptual channels. The reliability of a certain signature attribute ($\rho_A = \rho_{\gamma_k}$) is, like the attribute value itself in Subsec. IV-B2, calculated using only the compatible percept attributes.

Anchor reliability is a certainty measure of the statement that the anchor defines an object that is available in the real world with the properties described by the anchor signature. If the certainty of this statement is below a particular threshold, the *grounded threshold*, the hypothesis is not sustainable any more. These hypotheses are marked as *ungrounded*, but are still kept in the system until the reliability decreases below a further threshold. This approach is used to be able to re-identify objects that have not been perceived for a certain time, but are still present in the observed scenario. This can be the case due to short-time failure of sensory components or because the object is temporarily out of the sensor range.

C. Process weights

The weights and thus the quality of the perceptual processes are determined at two points in the overall system.

First, an a priori training step integrates scenario knowledge by separately assessing the quality of each perceptual process in this specific environment on the basis of annotated training data. This is done by

$$\tau^{prior} = \frac{pos_t}{pos_t + pos_f} \frac{obj_d}{obj_t}. \quad (11)$$

With pos_t and pos_f being true and false positives, obj_d and obj_t the number of detected and true objects, respectively. This weight is an adapted signal-to-noise ratio that accounts for the fact that more than one true positive can be counted for a truth object (assignment of multiple percepts to an hypothesis is possible), by integrating pos_t into the denominator.

Second, the process weights are adapted online to include changes in the quality of perceptual processes due to environmental changes such as varying lighting conditions. To accomplish this, a system-internal measure for process quality has to be defined. We assume that percepts that are generated by noise and do not refer to objects in the real world generate anchor hypotheses that do not get support by other perceptual processes or are even not strong enough to generate a hypothesis at all. Since these hypotheses are removed from the system in short time due to a very low reliability value, we can establish a connection between process quality and the time that percepts of this process have been assigned to an anchor hypothesis. The runtime process weight can thus be defined by

$$\tau_i^{run} = \frac{d_i^{mean}}{\sum_{k=0}^{K} d_k^{mean}} \quad (12)$$

with d_i^{mean} being the mean assignment time of percepts of type i, which can be calculated recursively. The overall weight τ is then calculated by combining the a priori and the runtime weight:

$$\tau = \frac{\tau^{prior} C^{prior} + \tau^{run} C^{run}}{C^{prior} + C^{run}} \quad (13)$$

The factor C^{run} is the total number of percepts received in the process. C^{prior} expresses the influence of the a priori training and should be set according to the quality of the training data and the expected environmental changes. It is crucial to note that the online adaption is no replacement for the training step because it works properly only if the weights have been initialized with adequate values.

V. Real-time Person tracking

To show the applicability of the generic anchoring approach for object tracking in a multi-process environment, we apply the anchoring system to the task of real-time person tracking. In this case, the different perceptual processes do not refer to sensors of different modalities but rather to video-based detection methods that produce different types of percepts.

A. Perceptual processes

We use four different *data-driven detection methods*: *motion detection* based on a temporal differencing, *foreground detection* that makes use of a reference image of the empty scene, an *object detection* process based on the work of P. Viola presented in [13], [14] and trained to detect *faces* and *person silhouettes*.

Due to inherent process characteristics, the motion caused by a single person cannot always be segmented in a single region and thus splits up into multiple percepts. Therefore, it should be allowed to assign those multiple percepts to a single anchor hypothesis, which leads to usage of the N:1 assignment method for motion percepts as well as for foreground percepts (due to the same process characteristics). Since the face and person percepts are inherently stable, we use the exclusive 1:1 assignment method for these.

Two different *expectation-driven processes* are integrated into the system.

A kernel-based *tracker* process can be started for a specific image region at runtime. This is done when certain conditions are fulfilled for an anchor hypothesis. In the special case of person tracking, we use three different reliability conditions: an anchor reliability condition, a reliability condition for the signature-position attribute and one for the signature-region attribute. Furthermore, a value-constraint condition for the region makes sure that a tracker process is started only for regions, the spatial extent of which does not differ too much from the person model that was acquired on training data. Once the tracker process has been triggered for an image region, it continuously searches for this region in input images until it is stopped. This is done by the anchoring system when the information that is provided by the tracker and the information that is present in the anchor hypothesis form a contrast. This can happen due to bad initialization of the tracker process or even failure in the tracker process itself.

A *color histogram* process provides a color histogram for a specific image region each time the process is triggered. This process is used only to update the color attribute inside the anchor signature and uses the same starting conditions as the tracker process.

All processes provide a *position* and a *region* attribute. Furthermore, the motion and foreground detection provide a *color* description by a histogram. This is used by the matching process and strengthens the ability to distinguish persons. The object detection process, which is trained on faces and person silhouettes, also provides a *class id* that distinguishes the object categories. This can be used in the matching process to avoid assignment to certain anchor types (which is not discussed in this paper).

All percepts except face are already generated in the correct reference system regarding the person center. The transfer function κ is thus applied only to face percepts. In this case, κ incorporates the offset of faces to the person center. Since the offset depends on the expected person size and the camera perspective, it is determined a priori on training data.

Table I
SIGNATURE ATTRIBUTES AND ACCORDING MATCHING METHODS

Attribute	Matching model
2D position	$\beta_{P2D}(\vec{p},\vec{x}) = \frac{\mathcal{N}_{\vec{x}}(\vec{p},K_{P2D})}{\mathcal{N}_{\vec{p}}(\vec{p},K_{P2D})}$
2D region	$\beta_{R2D}(\vec{r},\vec{x}) = \frac{\mathcal{N}_{\vec{x}}(\vec{r},K_{R2D})}{\mathcal{N}_{\vec{r}}(\vec{r},K_{R2D})}$
Class affiliation	$\beta_C(C_P,C_D) = \begin{cases} 1 & \text{if } C_P = C_D \\ 0 & \text{else} \end{cases}$
Color histogram	$\beta_H(H_P,H_A) = 1 - \sum_{i=1}^{I} \frac{(H_P(i)-H_A(i))^2}{(H_P(i)+H_A(i))}$

B. Anchoring of persons

1) Assignment: Table I shows an overview of the anchor signature attributes and the applied matching methods. These provide the attribute specific matching functions β_i used in equation (2). The similarity of two positions $\vec{p}$ and $\vec{x}$ is then defined by the value of the normal distribution with mean $\vec{p}$ and covariance K_{P2D} (which is determined in a training step), at position $\vec{x}$, normalized with the distribution's maximum value. The match of regions is determined equivalently. Color histograms are matched using a Chi-Square based similarity measure.

2) Update: The implementation of the *temporal fusion* is defined independently for each percept/attribute combination. We can thus use different fusion approaches for the same attribute types in different percepts. Since the percepts generated by the object detection and tracker processes are inherently stable statements by themselves, no temporal fusion is applied to them. An old percept of this type is completely replaced by a new one, which includes the replacement of the reliability value. Since the color histogram is acquired by a separate process that serves only to update the color attribute of the anchor signature, this histogram is also replaced by the newest value. In contrast to this, the position and region attributes in the motion and foreground detection processes are fused. In this fusion process, we must account for two characteristics. Multiple percepts can be assigned to a hypothesis at the same time and these percepts should be processed independently (which allows parallelism inside perceptual processes).

For the *position* attribute $\vec{x}$ of an incoming percept with timestamp T, this is done in two steps. First, $\vec{x}_{new}$ and ρ_{new} are calculated, whereby all percepts received in the same channel at the same time T contribute to the calculation.

$$\rho_{new} = \sum_{i=0}^{I} \rho_i \quad ; \quad \vec{x}_{new} = \sum_{i=0}^{I} \frac{\rho_i}{\rho_{new}} \vec{x}_i \tag{14}$$

Then, the new anchor position is calculated by

$$\vec{x} = \frac{\epsilon_{old}\rho_{old}(1-\rho_{new})\vec{x}_{old} + \rho_{new}\vec{x}_{new}}{\epsilon_{old}\rho_{old}(1-\rho_{new}) + \rho_{new}}. \tag{15}$$

The position is thus completely shifted to the new one only if the new percept has a reliability of 1 or the old percept is time-weighted with $\epsilon_{old} = 0$.

Since the *region* attribute is a description of the spatial dimension of a person, which is unlikely to vary significantly in a small time interval, it is desirable to retain high stability for this attribute over time. For this reason, the region is changed only if either the new percepts of a certain time do not fit the current region, in which case it is extended, or the reliabilities of the new percepts are high enough to assume that they represent a person correctly, in which case the region is adapted to fit the new percepts.

The *process fusion* follows the general scheme described in Subsec. IV-B2. To incorporate model knowledge into this process, the specific weighting factor ζ is used. For the *region* attribute $\vec{r}$, this weight incorporates the expected person aspect ratio $\vec{v}_{exp}$ that was determined on training data, i.e.,

$$\zeta_{R2D} = e^{-|\vec{v}_{exp} - \vec{v}_{\vec{r}}|}. \tag{16}$$

All other attributes use $\zeta = 1$.

C. Evaluation

The person tracking system is evaluated in the INRIA scenario of the CAVIAR data set [1], in a simulated real-time environment. Thus, the perceptual processes do not process each image in the video stream, but work with different frequencies that depend on the processing time of the underlying algorithms.

To assess the results of the anchoring system, the evaluation approach and the metrics introduced in [15] are used.

The first step of the application of the anchoring system in a specific scenario is the determination of the grounded threshold. This is important because this threshold defines which hypotheses are regarded as sufficient and which are not. To determine this value, different thresholds are evaluated on a training set. The grounded value is then chosen to be the one that maximizes the sum of multiple object tracker accuracy (MOTA) rates of the I training sequences:

$$\rho_g = \arg\max_{\rho} \left(\sum_{i=1}^{I} \text{MOTA}_i(\rho) \right) \tag{17}$$

The results of this process for the INRIA training sequences are shown in Fig. 3.

The video sequences in the INRIA scenario are subdivided into classes that involve different difficulties for a person tracking system. The evaluation of the anchoring system is done in a representative subset of sequences of each class. These results, as well as the total results of the INRIA scenario are shown in Table II. The table shows (in percent) the ratios of the multiple object tracker accuracy (MOTA), the misses ($\bar{m}$), the false positives ($\bar{f}p$) and the mismatches ($m\bar{m}e$). The multiple object tracker precision (MOTP) is shown in

[1] EC funded project CAVIAR (IST 2001 37540), see: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

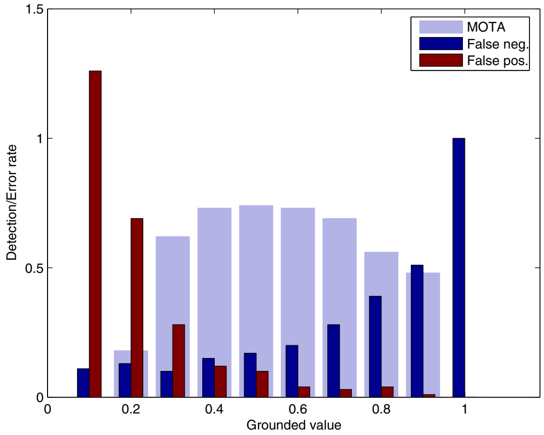

Figure 3. Determining the grounded value on a training set in the INRIA environment.

Table II
TRACKING RESULTS OF THE INRIA SCENARIOS.

Scenarios	MOTA	$\bar{m}$	$\bar{fp}$	$m\bar{m}e$	MOTP	Objects
Browsing	92.1	2.9	4.9	0.1	5.3	3 (807)
Leaving bags	91.2	3.4	5.4	0	4.2	2 (925)
Resting	81.3	15.6	3.1	0	6.2	3 (938)
Walking	80.2	12.4	7.3	0.1	7.7	5 (1749)
People groups	68.2	29.8	1.9	0.1	5.5	6 (773)
Fighting	54.2	29.1	16.2	0.5	7.5	7 (1438)
Total	72.1	20.1	7.8	0.1	6.3	4.4

pixels. The last column displays the number of different persons present in the video sequences and, in brackets, the total ground-truth person count of the used sequences (which reflects the amount of data used for evaluation).

The results show reasonable performance of the system in the first classes, where mainly single persons are present in the scenario. With the increasing number of persons in the scenario and especially the interaction between these persons, the results worsen. This is due mainly to the fact that the detection methods employed are not always sufficient to distinguish interacting persons or to detect individual persons in a group. To improve these results, additional data-driven perceptual processes could be integrated into the system. Furthermore, the problem of distinguishing persons could be solved by expectation-driven processes that verify existing anchor hypotheses in a top-down manner.

VI. CONCLUSION

In this paper, we proposed a generic method of combining perceptual processes for the task of object tracking. For this purpose, we provided an extension to the anchoring approach that permits application in multi-process environments. In this context, we introduced a two-step feature-level fusion strategy that is specially designed for application in heterogeneous-, asynchronous-, real-time-process environments. It was shown that the multi-modal anchoring approach can be successfully applied to the task of person tracking by combining multiple independent detection methods. The evaluation in this domain has also shown what difficulties arise in a generic approach, where the perceptual level is decoupled from the system. Our approach therefore permits the integration of runtime system knowledge into the perceptual level by expectation-driven processes. By this approach the decoupling of system and perceptual level can be relaxed without loosing the independence of the main system.

REFERENCES

[1] S. Coradeschi and A. Saffioti, “An Introduction to the Anchoring Problem”, in *Robotics and Autonomous Systems*, vol. 43, pp. 85–96, 2003.

[2] S. Coradeschi and A. Saffioti, “Anchoring Symbols to Sensor Data: Preliminary Report ”, in *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* , pp. 129-135, 2000.

[3] S. Lang, “Multimodale Aufmerksamkeitssteuerung für einen mobilen Roboter”,Phd. Thesis, Applied Computer Science, Bielefeld University, 2005.

[4] M. Kleinehagenbrock, S. Lang, J. Fritsch, F. Lömker, G. A. Fink and G. Sagerer, “Person Tracking with a Mobile Robot based on Multi-Modal Anchoring”, in *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication*, pp. 423-429, 2002.

[5] S. Lang, M. Kleinehagenbrock, J. Fritsch, G. A. Fink, and G. Sagerer, “Detection of Communication Partners from a Mobile Robot”, in *Proc. 4th Workshop on Dynamic Perception*, pp. 183-188, 2002.

[6] F. Klaus, “Einführung in Methoden und Techniken der Multisensor-Datenfusion”, PhD thesis, Universität Siegen, Fachbereich 12, Elektrotechnik und Informatik, 2003.

[7] M. Hanheide, “A Cognitive Ego-Vision System for Interactive Assistance”,Phd. Thesis, Applied Computer Science, Bielefeld University, 2006.

[8] S. Wachsmuth, S. Wrede, M. Hanheide, and C. Bauckhage, “An active memory model for cognitive computer vision systems”, in *KI-Journal, Special Issue on Cognitive Systems*, vol. 19(2), pp. 25-31, 2005.

[9] S. Wrede, J. Fritsch, C. Bauckhage, and G. Sagerer, “An xml based framework for cognitive vision architectures”, in *In Proc. Int. Conf. on Pattern Recognition, number 1*, pp. 757-760, 2004.

[10] S. Wrede, M. Hanheide, C. Bauckhage, and G. Sagerer, “An active memory as a model for information fusion”, in *In Int. Conf. on Information Fusion, number 1*, pp. 198-205, 2004.

[11] J. Munkres, “Algorithms for the Assignment and Transportation Problems”, in *Journal of the Society of Industrial and Applied Mathematics*, pp. 32-38, 1957.

[12] H. W. Kuhn, “The Hungarian Method for the assignment problem”, in *Naval Research Logistic Quarterly*, pp. 83-97, 1955.

[13] P. Viola and M. Jones, “Fast and robust classification using asymmetric AdaBoost and a detector cascade”, 2001.

[14] P. Viola and M. Jones, “Robust Real-time Object Detection”,in *Proc. IEEE Workshop on Statistical and Theories of Computer Vision*, 2001.

[15] K. Bernardi, A. Elbs and R. Stiefelhagen, “Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment”, *The Sixth IEEE International Workshop on Visual Surveillance*, 2006.

[16] J. Llinas and D. L. Hall., “Handbook of Multisensor Data Fusion”, *CRC Press*, 2001.

[17] E. Tulving, “Organization of memory: quo vadis?”, in *The Cognitive Neurosciences, MIT Press*, pp. 839-847, 1995.