Qualitative Abstraction and Inherent Uncertainty in Scene Recognition

Carsten Elfers, Otthein Herzog, Andrea Miene, and Thomas Wagner

Center for Computing Technologies (TZI), Universität Bremen, D-28359 Bremen
{celfers, herzog, miene, twagner}@tzi.de

1 Introduction

Scene interpretation has been identified as one of the most fundamental and challenging tasks already since Shakey [7]. In recent years the problem has often been addressed independently with varying foci within AI, vision, and robotics. But new ambitious (benchmark) tasks like ambient intelligence and *service- and entertainment robotics* (e.g., RoboCup) had shown the necessity of integration of formerly independent solutions. This integration process requires solutions to some new inherent difficulties: E.g., localization is one of the most fundamental tasks for physically grounded robots that also constitutes the map generation of the surrounding environment (e.g., by SLAM approaches [3], [5],[16]). The currently successful methods (i.e., Monte-Carlo, (extended) Kalman-Filter see e.g., [17]) handle sensor noise based on a set of probabilistic methods which directly result in probabilistic representations of the environment. In contrast, most classic knowledge representation techniques are based on strict non-probabilistic representations.

Generally, the resulting problem can be addressed in at least three different ways:

1. Minimizing uncertainty and (trying to) generate a strictly declarative representation with full inference power.
2. Handling uncertainty by the generation of a hybrid representation which incorporates probabilistic and declarative representations at the expense of less powerful declarative inferences.
3. Directly represent uncertainty by the use of a strictly probabilistic representation.

Each approach imposes strengths as well as (strict) limitations on the inferences available for scene interpretation. While deductive reasoning (e.g., subsumption) strongly relies on the first (or at least on the second) approach other inferences like *filtering*, *prediction* and *smoothing* rely on the latter representation. Nevertheless, scene interpretation requires both types of inference.

In this paper we present two different approaches that allow us to incorporate both types of inferences. Generally, continuous quantitative sensor information can be abstracted in terms of qualitative representations in two different ways:

1. focusing on strictly discrete percepts (while ignoring continuous information) and
2. qualitative abstraction into predefined frames of references.

In the first part of this paper (in section 2) we present an approach that combines the focus on discrete percepts and the generation of qualitative abstractions[1]. The localization task is solved by a strict use of (qualitative) ordering information while the overall dynamic and static spatial world model is generated by qualitative abstraction. This overall process results in a non-probabilistic representation.

Nevertheless, uncertainty in scene interpretation arises also independently from specific sensor noise due to either missing information or ambiguity in interpretation. Therefore, in the second part of the paper (in section 3) we present an approach to prediction that allows for the flexible and precise prediction of actions at different levels of granularity in the scene-specific required precision and the data available.

[1] This part strongly relies on [12].

Dagstuhl Seminar Proceedings 08091
Logic and Probability for Scene Interpretation
http://drops.dagstuhl.de/opus/volltexte/2008/1614

2 Generation of qualitative spatial representations

2.1 Motivation and Related Work

The generation of qualitative ordinal knowledge and its use for qualitative navigation has been investigated practically as well as theoretically. The idea has been introduced by Levitt and Lawton as part of their *QUALNAV*-approach [8]. Imagine walking through an unknown city during a conference visit. You see different landmarks: a large office building far away on your left, a church on your right and a large railway station in your back. The underlying hypothesis of Levitt and Lawton is that the full 360^0 ordering (*roundview*) in which a set of landmarks is perceived by some omnidirectional sensor of an autonomous systems is directly related to the specific position of the observer. Or the other way a round, the position of the observer is directly related to the ordering in which a set of landmarks is perceived. Although the idea appears to be intuitive when we consider our own experience of landmark use walking through an unknown city the hypothesis of Lewitt and Lawton does not hold in general. The example in figure 1 shows a simple counter-example (adopted from Schlieder [15]). The position of the autonomous system is indicated as a black dot. Due to Levitt and Lawton each region which results from connecting each landmark with each other (i.e., an *arrangement*) should be identified by a specific ordering. In picture 1 the cyclic ordering is given by $\langle 1, 2, 3, 4, 5 \rangle$.

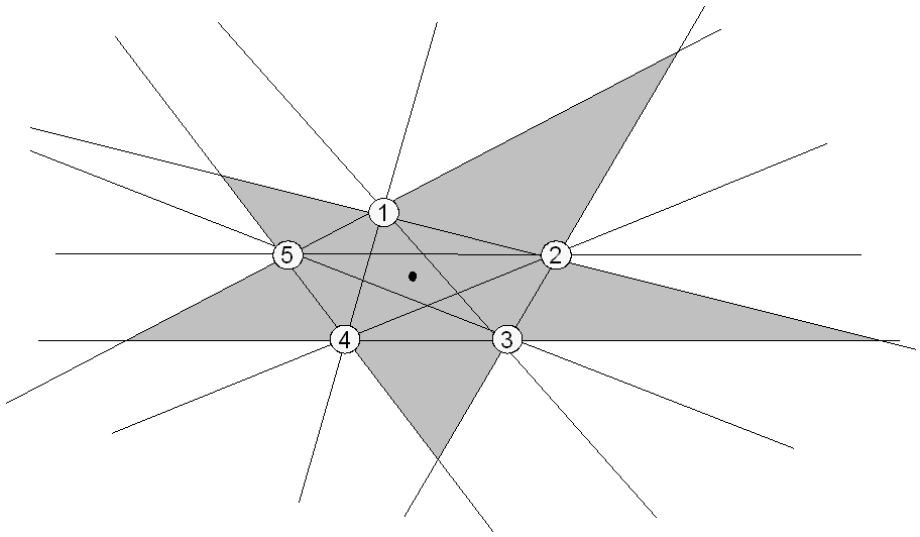


Fig. 1. Localization and ordering information

But the resulting circular ordering information is not unique to a specific region is valid instead for all grey regions.

The detailed formal analysis of Schlieder ([15], [13], [14]) showed that the information encoded in the *roundview* of Levitt and Lawton is not sufficient for qualitative navigation/localization. Schlieder proposed instead an extended *panoramic* representation that incorporates the opposite sides of landmarks for which he could proof a bijective mapping between qualitative position and landmark ordering. For practical applications the information requirements are very high. We do not only need a full 360^0 view but we also have to incorporate the opposite landmark sides which by definition cannot be perceived directly and therefore have to be calculated (e.g., based on angular information).

In section 2.2 we present a view-based approach to qualitative navigation that requires only partial egocentric views (i.e., neither 360^0 views nor opposite landmark sides) but still allows for a robust mapping between position and perception.

Another field of application for qualitative knowledge is the qualitative description of motion and the correlated interpretation and prediction of dynamic scenes. Dynamic scenes consist of objects which are in certain spatial relations to each other. The relations vary over time due to the movement of the objects. Temporal intervals and the relations between them can be represented following the approaches of Allen [1] and Freksa [6]. The spatial relations between the objects can be described using metric knowledge such as angles, distances and the objects movement in terms of direction and speed. Quantitative values concerning distances and directions can be mapped onto qualitative classes using qualitative distance measures as proposed by Hernandez [4]. For a detailed discussion of the related work please refer to [9]. An approach which brings together the temporal and spatial aspects to describe, interpret and predict dynamic scenes is presented in section 2.3.

2.2 Qualitative Localization Based on Egocentric Views

Localization and navigation can be interpreted as the mapping between perception and space. In the case of the traditional approaches [18] the Euclidian 2D/3D space is used as the reference system and perception is given in terms of quantitative sensor output. In the case of qualitative localization both perception and the spatial reference system must be defined with respect to some qualitative reference system. In the following sections the concept of view-based qualitative navigation is demonstrated with landmarks configurations with four landmarks, although the general concept is not limited to any specific number of landmarks. (For a full description please refer to [20].), Therefore we have to give,

1. a definition of the construction of qualitative perception,
2. the specification of a qualitative reference system and
3. the mapping from perception to space (localization)[2].

The fundamental idea of view-based navigation is to use the egocentric perception of an agent without a mapping into any allocentric reference system. The only information used to describe perception is ordering information, i.e., no angular nor any distance information will be used. Usually the generation of spatial qualitative descriptions is a difficult task due to the required classification process. In the case of ordering information the generation does not require any kind of classification. The idea is to fix an arbitrary point within the convex hull of a landmark configuration. The ordering information is given by the orthogonal projection of the landmarks on $L_{Orth(P_\Gamma/VP)}$ (see also figure 2). Formally[3],

Definition 1: (Snapshot Generation) Let P_Γ denote the position of an agent A_Γ and $C_{P(ABCD)}$ the parallelogram configuration formed by the set of points A, B, C, D in the plane. The line $L_{P_\Gamma/VP}$ is the line of vision from P_Γ to *VP*, with *VP* being a fixed point within $C_{P(ABCD)}$. Furthermore $L_{Orth(P_\Gamma/VP)}$ be the orthogonal intersection of $L_{P_\Gamma/VP}$. The landmark panoramic ordering information can then be described by the orthogonal projection $P(P_\Gamma, VP, C_{P(ABCD)})$ of the points $ABCD$ onto $L_{Orth(P_\Gamma/VP)}$.

Assume a parallelogram configuration $C_{P(ABCD)}$ of the landmarks $A, B, C, D \in \mathcal{L}$ with all landmarks connected to each other by a straight line $L_{n/m}$, $n, m \in \mathcal{L}$[4]. The resulting structure decomposes the space in twelve region outside the convex hull of $C_{P(ABCD)}$.

[2] The crucial point is to show that there is a bijective mapping between perception and qualitative position. This can be shown e.g., by constructing an appropriate *finite state machine*. For details please refer to [20].

[3] The generation of a complete ordinal snapshot as described in definition 1 is only necessary for the initial construction of the reference system.

[4] There are no specific requirements for parallelogram configuration e.g., a rectangle.

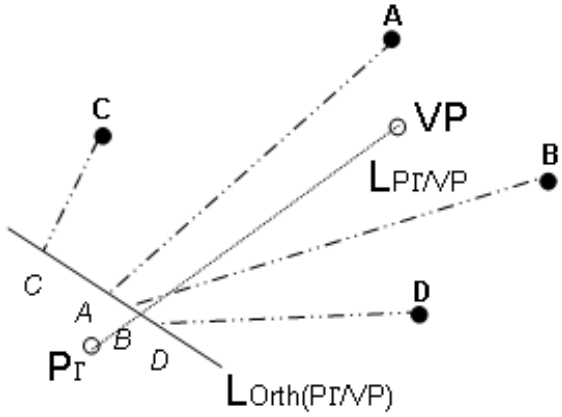


Fig. 2. Construction of an ordering view

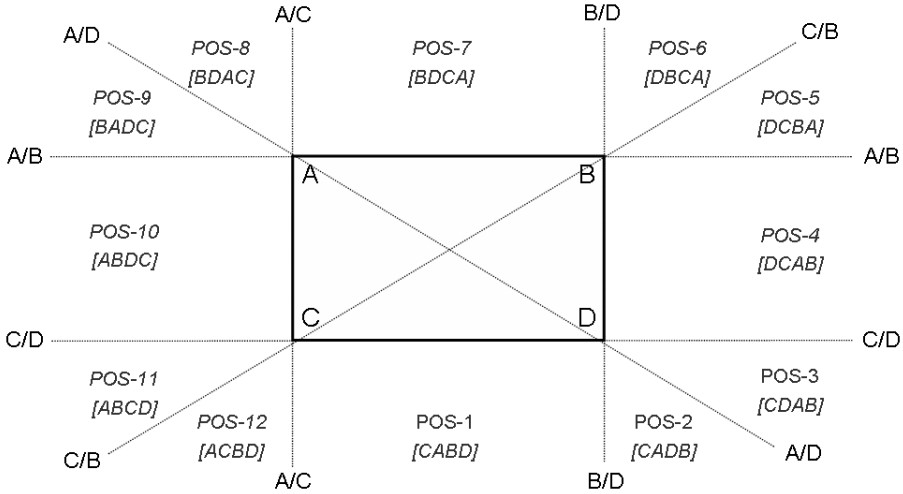


Fig. 3. Qualitative regions, transitions und ordinal perceptions for a parallelogram landmark configuration

Moving around $C_{P(ABCD)}$ either clockwise or counter-clockwise results in a set of ordering snapshots that describe qualitatively the position of the observer with respect to $C_{P(ABCD)}$,

Observation 1: (Parallelogram Snapshot Cycle) The panoramic landmark representations resulting from the subsequent projection $P(P_\Gamma, VP, C_{P(ABCD)})$ by counter-clockwise circular movement around *VP* can be described by the following ordered, circular sequence of snapshots: $((ABCD)$,$(ACBD)$, $(CABD)$, $(CADB)$, $(CDAB)$, $(DCBA)$, $(DCAB)$, $(DBCA)$, $(BDCA)$, $(BDAC)$,
$(BADC)$, $(ABDC))$

Each line $L_{n/m}$ connecting landmarks n and m with each other can be interpreted as a transition axis. Given the agent is located at position $[POS-7]$ with the associated perception $\langle BDCA \rangle$ and is moving counterclockwise towards region $[POS-8]$. While passing the transitions axis $L_{A/C}$ the ordering perception changes from $\langle BDCA \rangle$ to $\langle BDAC \rangle$, Considering the result of a full round walk the ordering topology of observation 1 can alternatively be described in terms of a sequence of transitions: $\langle$ B/D, A/D,

C/D, A/B, C/B, B/D, A/C, A/D, A/B, C/D, C/B, A/C $\rangle$[5]. During the navigation around $C_{P(ABCD)}$ each transition axis $L_{n/m}$ is passed exactly twice. Thus the observation of a transition is at least to some extent invariant. But the navigation process is even more constrained. Given the transition axis $L_{n/m}$ we are able to distinguish on which side a robot passes $L_{n/m}$. In case of $L_{A/C}$ the landmark A moves from the right to the left and landmark C moves from the left to the right side in the case of moving from region $[POS-7]$ to $[POS-8]$. While passing the $L_{A/C}$ on the bottom side from region $[POS-12]$ to $[POS-1]$ (once again, we assume a counter-clockwise direction of navigation) the landmark switch is exactly the other way a round. So the navigation can be described more precisely as,$\langle$ C/D, A/B, C/B, D/B, A/C, D/A, B/A, D/C, B/C, C/A $\rangle$. The fundamental advantage of describing a landmark configuration in terms of a transition sequence is that only a minimum of information is required to determine the observers, positions. Just observing, e.g., the landmark switch A/C in combination with the direction of navigation (clockwise vs. counterclockwise), the direction of the landmark switch allows to determine the exact observer position with respect to $C_{P(ABCD)}$.

An additional interesting feature of ordering information is that it is, e.g., variant to various deformations like compression. The circular sequence of snapshots described in *observation 1* is indeed only valid for parallelogram landmark configurations. Imagine we are moving the landmarks B and D in figure 3 towards each other. As a matter of consequence the transitions axis $L_{A/B}$ and $L_{C/D}$ have no longer a parallel orientation. Instead after moving the landmarks B and D towards each other the axis $L_{A/B}$ and $L_{C/D}$ will intersect on the right side of $C_{P(ABCD)}$ and create a new region $\langle CDBA \rangle$. Generally four new regions may arise depending on which landmarks are changing their relativ position to each other. This allows us to describe the second observation,

Observation 2: A semi-irregular formed quad-tuple configuration, i.e., with two parallel lines either L_{AC} and $L_{B/D}$ or $L_{A/B}$ and $L_{C/D}$, will generate the following additional state:

$$((DBAC) \vee_{XOR} (ACDB)) \vee_{XOR} ((BACD) \vee_{XOR} (CDBA))$$

The new positions cannot be combined arbitrarily. Lets assume the same case as above. The landmarks B and D are moved towards each other and therefore the axis $L_{A/B}$ and $L_{C/D}$ will intersect on the right side of $C_{P(ABCD)}$. Since no straight lines, i.e., $L_{A/B}$ and $L_{C/D}$, can intersect more than once it is clear that $L_{A/B}$ and $L_{C/D}$ will not intersect on the left side of $C_{P(ABCD)}$. Thus any landmark configuration with four points has at most two additional regions (in addition to the ones specified in *observation 1*),

Observation 3: A irregular formed quad-tuple configuration, i.e., with no parallel lines $L_{A/C}$, $L_{B/D}$, $L_{A/B}$ and $L_{C/D}$, will generate the following additional states:

$$((DBAC) \vee_{XOR} (ACDB)) \wedge ((BACD) \vee_{XOR} (CDBA))$$

Thus we are able to distinguish nine different convex quad-tuple configurations by a strict analysis of the ordering snapshots (see figure 4).

The approach has been tested in two different scenarios. First it was tested in the *RoboCup* domain with a simulator of the *Sony-Four-Legged-League* [19]. Since our approach is intended to be used for localization *outside* the convex hull of a landmark configuration the edges of the lines *within* the soccer field were used as landmarks (for detailed description please refer to [19]). Secondly, in order to get results that do not depend on any specific kind of landmark (re-)detection we also developed a simulator *EGO-QUALNAV* that allows to control precisely various fault modes like odometrie, missing landmarks,

[5] Therefore, the *parallelogram snapshot cycle* (Observation 1) does not require to focus on some arbitrary viewpoint VP. Instead the observation of the transitions is sufficient. The point VP of definition 1 is only required for the initial reference view.

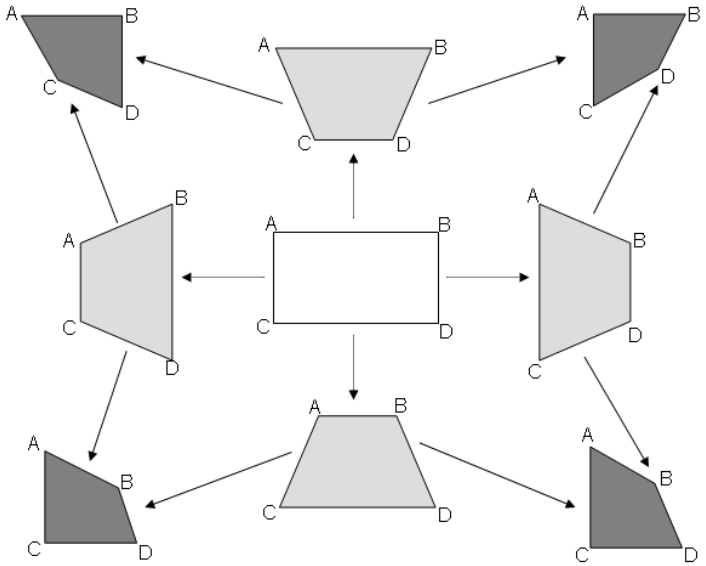
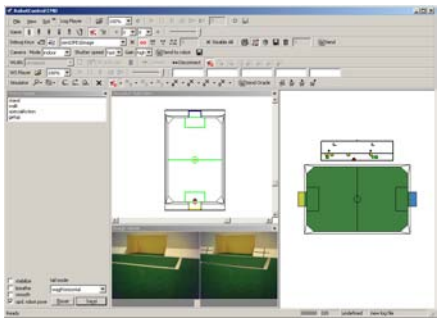
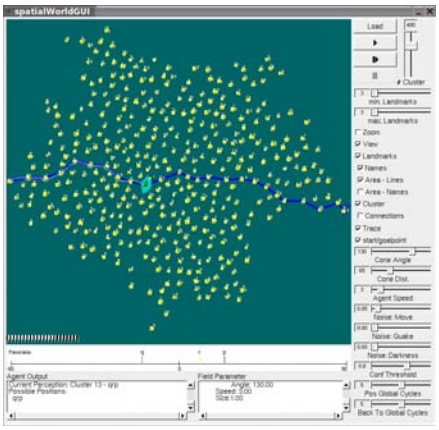


Fig. 4. Convex quad-tuple ordering topologies

partial views and wrong identification of landmarks. (Figure 5(b) shows a more complex scenario. Each bright dot describes a landmark configuration (a cluster) whereas all landmark configurations are connected to each other by an *accessibility* relation in order to construct more complex scenarios.)



(a) Validation in the RoboCup-domain (Simulator of the *Sony-Four-Legged-League*)



(b) EGO-QUALNAV-SIM - environment with a graph-based network of landmark configurations

Fig. 5. Validation

Even in cases where up to 60% of the perception is incorrect[6] and with a high rate of missing information (e.g., landmarks that could not be distinguished) the simulated agent was able to find its way from an arbitrary starting point to an arbitrary end point (for a detailed description of the results and the precise formalisation with the according proofs please refer to [20]).

[6] Each test case has been tested with 10000 navigation tasks. For detailed results please refer to [20].

2.3 Dynamic Qualitative Information

In this section we introduce our approach on representing motion with qualitative dynamic knowledge. The approach enables us to both interpret and predict complex dynamic situations [9], [11].

Qualitative Motion Description The description includes single object's motion in combination with the changes in the objects' pairwise spatial relations over time. The basic assumption of our approach is that we have an allocentric view from above of the motion scene. On a quantitative level the objects, absolute and relative movement is described by four types of time series: the motion direction and speed of each object, and the spatial direction and distance for each pair of objects. In a first abstraction step each time series is segmented into time intervals of homogeneous motion values.

In order to segment the time series into time intervals two different segmentation methods are used: a threshold-based segmentation method and a monotonicity-based segmentation method, which groups together strictly monotonic increasing intervals, strictly monotonic decreasing intervals and intervals of constant values. Each threshold-based segmented interval is described by a single attribute: the average of its values. A monotonicity-based segmented interval is described by its start value, its end value, and the run direction of values: increasing, decreasing or constant. Both segmentation methods allow for various interpretations of the resulting intervals. The monotonicity-based segmentation is useful to recognize dynamic aspects of motion, e.g., the acceleration of a moving object. But due to the fact that the values are measured only at the start and the end of an interval its intermediate values are not known. Therefore, the threshold-based segmentation is more useful to find, e.g. an object that moves with a certain average speed. In a second step the attribute values describing the intervals are mapped onto qualitative classes for direction, speed or distance, respectively using qualitative distance measures as suggested by Hernandez [4]. The entire process is carried out online, i.e., at each time cycle one set of positional data is processed. Fig. 6 shows the entire process of motion description exemplary for a time series of object distances, segmented using the monotonicity-based method. A single interval already allows for a simple interpretation of the movement of the two involved objects: they approach each other and finally meet, which is expressed by the term HOLDS(approach-and-meet$(p, q), \langle t_n, t_{n+k} \rangle$). The predicate HOLDS expresses the coherence between a certain situation and the time interval in which it is taking place or is valid (see Allen [1]).

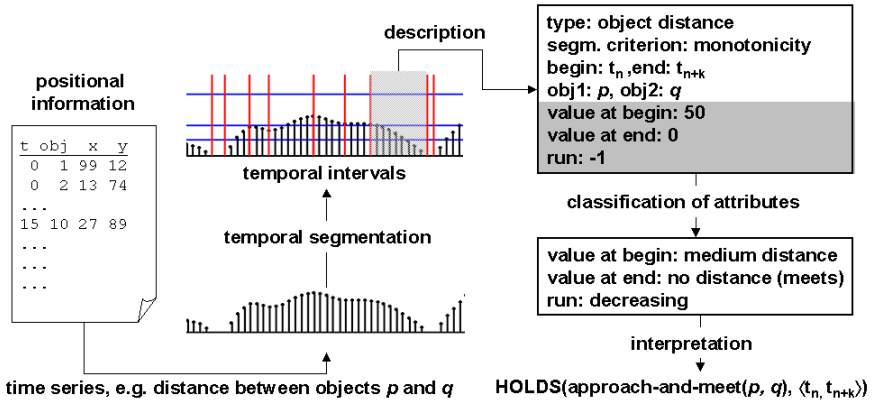


Fig. 6. Overview: Generation of motion description

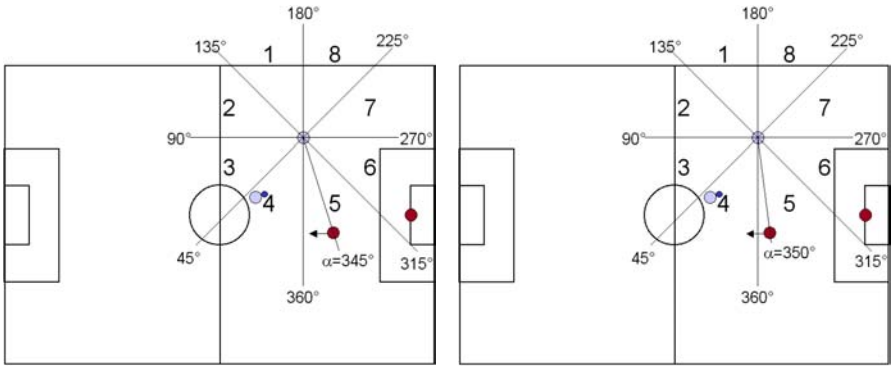


Fig. 7. Development of spatial directions between offender and defender announcing an impending offside position.

Interpretation and Prediction of Dynamic Scenes Based on the qualitative motion description it is possible to recognize and predict motion situations. Domain knowledge, e.g. about the function or type of objects involved in a situation, leads to more appropriate interpretations. In addition, positional information is integrated by representing the duration a certain object is located in a certain region via time intervals.

As an example it is possible to predict an impending offside trap (FIFA rules, law 11). In order to predict an impending offside situation for player p, he has to be located in the opponents' half, actually have the ball behind him and a small remaining number of $k = 3 - 4$ opponent defenders in front of him. Then it depends on the relative movement of p and an opponent q if an offside position is impending. Therefore, we have to take into account the current spatial direction between p and q (spatdir), obtained from the threshold-based segmentation, and the development of the spatial direction between p and q (clockwise (change-spatdir-cw) or counterclockwise (change-spatdir-ccw), obtained from the monotonicity-based segmentation). If the spatial direction is already close to the change between in-front-of and behind, and the values are increasing or decreasing (clockwise/counterclockwise change of spatial directions) an offside position is impending.

$$
\begin{aligned}
&\text{HOLDS}(\textsf{offside-danger}(p,q), \langle max(s_i), min(e_i)\rangle) \Leftrightarrow \\
&\exists \langle s_i, e_i\rangle, i \in \{1,\ldots,6\} : \\
&\quad \text{HOLDS}(\textsf{region}(p, \textit{opponent-half}), \langle s_1, e_1\rangle) \wedge \\
&\quad \text{HOLDS}(\textsf{behind}(ball, p), \langle s_2, e_2\rangle) \wedge \\
&\quad \text{HOLDS}(\textsf{in-front-of}(q,p), \langle s_3, e_3\rangle) \wedge \textsf{team}(p) \neq \textsf{team}(q) \wedge \\
&\quad \text{HOLDS}(\textsf{number-of-opponents-in-front-of}(p,n), \langle s_4, e_4\rangle) \wedge 2 \leq n < k \wedge \\
&\quad ((\text{HOLDS}(\textsf{change-spatdir-cw}(p,q), \langle s_5, e_5\rangle) \wedge \\
&\quad\ \ \text{HOLDS}(\textsf{spatdir}(p,q,1 \vee 5), \langle s_6, e_6\rangle)) \vee \\
&\quad\ (\text{HOLDS}(\textsf{change-spatdir-ccw}(p,q), \langle s_5, e_5\rangle) \wedge \\
&\quad\ \ \text{HOLDS}(\textsf{spatdir}(p,q,4 \vee 8), \langle s_6, e_6\rangle))) \wedge \\
&\quad \forall i,j \in \{1,\ldots,6\} : s_i < e_j.
\end{aligned}
$$

A complex situation like offside-danger(p, q) combines several time intervals. The temporal relations between the intervals are modelled using temporal relations on time intervals defined by Allen [1] and on semi-intervals as proposed by Freksa [6]. The term $\forall i,j \in \{1,\ldots,n\} : s_i < e_j$ postulates that all n intervals involved in the situation are pairwise contemporary. $\langle max(s_i), min(e_i)\rangle$ specifies the sub-interval covered by all n time intervals $\langle s_i, e_i\rangle, 1 \leq i \leq n$. Fig. 7 shows the case of an increasing development of values. If the present trend lasts for some further time, an offside situation will occur in the moment the spatial relation changes to the next class (i.e. from 5 to 4) and at the same point in time from in-front-of to behind.

Within the prediction phase we can also distinguish offside traps caused by a forward movement of an opponent q from offside situations caused solely by the movement of the offender p himself be taking into account the movement of these players.

To evaluate our approach we have chosen three games from the Robocup Worldcup 2002: FC Portugal vs. Puppets, TsinghuAeolus vs. FC Portugal and VW2002 vs. Cyberoos. The games include 53 offside situations in which the game was interrupted by the referee. In 45 cases our system also detected an offside situation. In 8 situations our systems is not in line with the referee. But in all of these situations the referee decides offside against a team A although a player of team B has touched the ball before the game was interrupted. So our system detected every correct offside situation and furthermore 8 wrong decisions of the referee.

A detailed explanation of the offside example together with the in-depth evaluation of results is presented in [11].

3 Relational Hidden Markov Model: Hierarchical Prediction based on Sensor Information

3.1 Learning and Prediction based on RHMM

Instead of creating a complex model like DPRMs, the model in this work is limited to a HMM similar structure to provide fast inference without the need of approximation methods. It also uses relational features to improve inference accuracy and to handle sparse reference data.

Relational Hidden Markov Model

To provide explicit modeling of sensory uncertainty and the ability to deal with sparse reference data the *relational Markov model* and the *hidden Markov model* have been combined in this work. The proposed method also provides inferences on different granularity levels. Like a *HMM* the *RHMM* is separated in hidden and visible states, and like a *RMM* each state is represented by a relation.

Definition 1. *The* relational hidden Markov model *is defined as a tuple* $RHMM =< \mathcal{D}, \mathcal{R}, \mathcal{E}, A, B, \pi >$ *with the set of all domains* $\mathcal{D}$*, the set of all hidden relations* $\mathcal{R}$*, the set of all visible relations* $\mathcal{W}$*, the transitionmatrix* A*, the sensormatrix* B *and the initial distribution* π*.*

To provide a detailed overview of the *RHMM* the special structure of it will be introduced in the next section, before going into detail about the inferences.

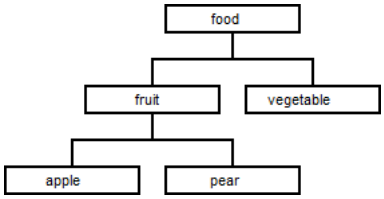


Fig. 8. Domain draft

Structure. To describe the domain-specific similarities we define the hidden states of the *RHMM* as a set of relations $\mathcal{R}$ and the visible states (evidences) as a set of relations $\mathcal{E}$, with each relation containing a set of attributes out of $\mathcal{A}$. The attribute values and the similarities between them are specified by a set of domains $\mathcal{D}$, one domain $D \in \mathcal{D}$ for each attribute. Therefore we define a domain as a hierarchical structure specifying the different granularity levels. A similarity between different values of an attribute is expressed by combining them into one value on a more abstract granularity level in the domain. An example for the structure of a domain is exemplified in figure 8. The visible and hidden states are handled in the same way, so we further omit defining both cases. To define a set of relations we specify a function called $leaves(\delta)$, gathering all leaf nodes

from a given node δ in the corresponding domain. With this function we define a relation $R(d_1, \cdots, d_k)$ by its containing ground relations as: [2]

$$R(d_1, \cdots, d_k) = \{R(\delta_1, \cdots, \delta_k) \in \mathcal{S} | \delta_i \in leaves(d_i). \\ \forall i (1 \leq i \leq k)\} \tag{1}$$

To specify how abstractions of predicates are built from the domains we omitted using all possible abstractions (like in *RMMs*) due to a high computational effort. Instead we build an abstraction by abstracting all attributes at the same time. Therefore we define a function $depth(d)$ returning the depth of a node in a domain from its most abstract root. For example the depth of the most abstract value of a domain is zero. Further the boolean function $min(d_1, d_2)$ is only fulfilled if the difference between the *depth* of the parameters is minimal in the corresponding domains. The abstraction $\mathcal{G}(s)$ of a ground predicate $s = R(\delta_1, \cdots, \delta_k)$ is defined as:

$$\mathcal{G}(s) = \{R(d_1, \cdots, d_k) \subseteq \mathcal{R} | d_i \in nodes(D_i) \\ \wedge \delta_i \in leaves(d_i) \wedge min(depth(d_i), depth(d_j)). \\ \forall i \forall j (1 \leq i, j \leq k\} \tag{2}$$

Inference Basically the inference in a *RHMM* is a combination of the inference in *RMMs* and the inference in *HMMs*. To determine the probability of a state transition $a_{i,j}$ we consider all more abstract state transitions $a_{\alpha,\beta}$ of the requested one like in *RMMs*. α and β therefore specify a relation on a more abstract granularity level:

$$a_{i,j} = P(q_t = S_i | q_{t-1} = S_j) \\ = \sum_{\alpha \in \mathcal{G}(q_t)} \sum_{\beta \in \mathcal{G}(q_{t-1})} \lambda_{\alpha,\beta} a_{\alpha,\beta} P(q_t | \beta) \tag{3}$$

Therefore $a_{\alpha,\beta}$ determines the transition probability of a more abstract state transition by including all containing transition probabilities in the calculation of the given transition probability as follows:

$$a_{\alpha,\beta} = \sum_{s_i \in \alpha} P(s_i | \alpha) \sum_{s_j \in \beta} o_{i,j} \tag{4}$$

$o_{i,j}$ represents the original trained state transition probability. To include more similar state transitions stronger than less similar state transition we used the proposed mixturefunction of the *RMM-Rank* method.

$$\lambda_{\alpha,\beta} \propto (\frac{n_{\alpha,\beta}}{10})^{rank(\alpha)+rank(\beta)} \tag{5}$$

The rank function is defined as $rank(R(d_1, \cdots, d_k) = 1 + \sum_{i=1}^{k} depth(d_i)$. Lambda is chosen that $\sum_{\alpha,\beta} \lambda_{\alpha,\beta} = 1$. $n_{\alpha,\beta}$ is the amount of state transitions from a predicate α to a predicate β. Analogical to $a_{i,j}$ we determine the emission probabilities $b_{i,j}$ on a different set of predicates, domains and attributes. For inference in the *RHMM* the *FORWARD-Algorithm* known from *HMMs* is used for inferences in *RHMMs* too:

$$P(Q_{t+1} | e_{1:t+1}) = \\ \alpha P(e_{t+1} | Q_{t+1}) \sum_{q_t \in Q_t} P(Q_{t+1} | q_t) P(q_t | e_{1:t}) \tag{6}$$

where α is a factor ensuring that the resulting state distribution sums up to one. $P(e_{t+1}|Q_{t+1})$ represents the sensor model and is determined by $b_{i,j}$. $P(Q_{t+1}|q_t)$ represents the transition model and is determined by $a_{i,j}$.
To approximatively compute the probability of a state/relation on a higher granularity level after inference the containing states on the lowest granularity level can simply be combined using the following equation:

$$P(R_t \in \mathcal{R}) = \sum_{s_i \in R_t} P(s_i) \tag{7}$$

Like in HMMs it is necessary that the states on all granularity levels are disjunct.

Training We assumed hidden but not invisible states to perform a simple maximum likelihood estimation for training the *RHMM* (s. Eqn. 8). Therefore the model will be trained by determining the relative frequency of the state transitions and recognized evidences in the given states.

$$a_{i,j} = \frac{\sum_t N(S_{t+1} = j, S_t = i)}{\sum_t N(S_t = i)} \tag{8}$$

N counts the state transitions of the parameters.
Determining the model parameters without knowledge about the hidden states the *Baum-Welch* method is applicable but depending on the specified model structure it may be very complex. Therefore N guesses the amount of state transitions instead of counting them. In contrast to the ML method, the *Baum-Welch* method is an iterative process where the state transition probabilities have to be determined after each iteration which may result in a high computational effort.

Representing the RoboCup domain

To represent the essential features of the environment, two attribute domains have been specified: The distance and the direction to represent relative coordinates. Figure 9 illustrates the distance and direction domains, e.g., the distance domain with four separate values on the finest granularity level. On the next more abstract level these four states are combined to two states, e.g., the distance values $Near$ and $Middle$ are combined to the value $AnyNear$.

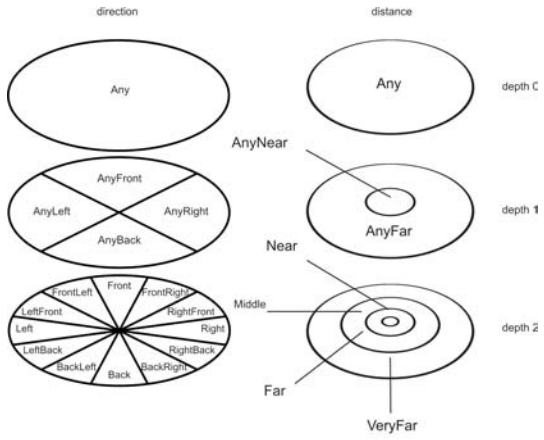


Fig. 9. Illustration of the distance and direction domains.

The direction domain is specified like the distance domain on three separate levels of granularity. On the finest granularity level 12 separate states are distinguished. Additionally each attribute domain gets a value $none$ on the finest granularity level and the value $anyNone$ containing $none$ on the second finest granularity level connected to the root (Any) to offer a state to represent features that cannot be determined, e.g., unseen objects.
For the hidden states we created two predicates, $dribble(direction, distance)$ and $pass(direction, distance)$ leading to 130 hypotheses on the finest granularity level, 30 hypotheses on the next granularity level and two hypotheses on the most abstract granularity level to distinguish. For the visible states we experimented with different features of the environment, e.g., the relative

position of the nearest teammate, the relative position of the next opponent and more. A simple heuristical method to determine the best position on the field regarding the distance to the goal, the distance to the ball owner, a negative influence of near opponent players and a positive influence of near teammates turned out to be most suitable after a short period of tests. The visible states therefore represent the relative position to the best heuristically determined position on the field. Therefore one predicate represents the evidence, $evidence(direction, distance)$, with 65 states to distinguish.

After the discretisation of the states the time also needs to be discretised. Therefore the actions build dynamic time frames. Each action is recognized by a symbolic method based on [9]. The evidence will be perceived after an action ends respectively each time an action starts. Based on these evidence/action pairs the model will be trained and predict the following actions based on the evidences.

3.2 Evaluation

The *RHMM* has been evaluated in the *Simulation League 3D RoboCupSoccer* domain. The necessary reference data has been generated offline by a symbolic action recognition tool, based on a method from *A. Miene* [10]. Therefore 20 games of the team *Virtual Werder 3D* have been recorded. By the fact that the reference data is offline available a leave-one-out cross-validation has been performed to ensure a high degree of accuracy.

Comparable Results First we tested how the *RHMM* assesses the 130 different hypotheses depending on the amount of available training sequences (354 Sequences have been gathered during the 20 games). To generate comparable results we performed this test with exactly the same data with the *HMM* and a simple symbolic method (called *SIMPLE*) setting the probability of the hypothesis to 100% if it occured in the trained data, 0% if not. For all these three methods we measured the amount of wrong hypotheses assessed higher than the right one.

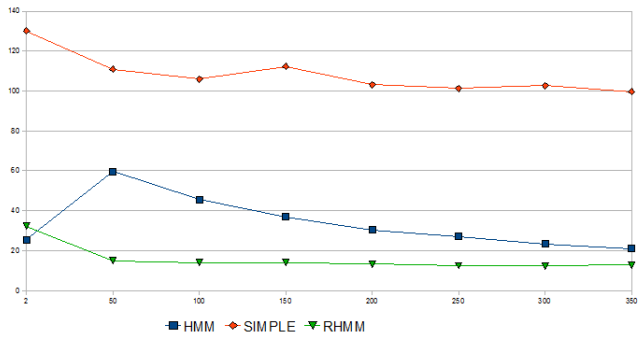


Fig. 10. Hypothesis rank test

Figure 10 shows the results indicating that the *RHMM* predominantly outperforms the *HMM* and *SIMPLE* method. Especially the *SIMPLE* method seems to be an inappropriate method for action prediction in such an uncertain environment. The simple method predicted the right action in an average after $107,32$ wrong hypotheses. Basically the inability of representing beliefs of certain evidences seems to be responsible for the enormous amount of errors. Also the *RHMM* predominantly outperforms the traditional *HMM* with an average error of $13,74$ to $37,05$ in evaluating the hypotheses, especially with a low

amount of reference data. This shows that considering domain-specific information during computation influences the result in a positive way.

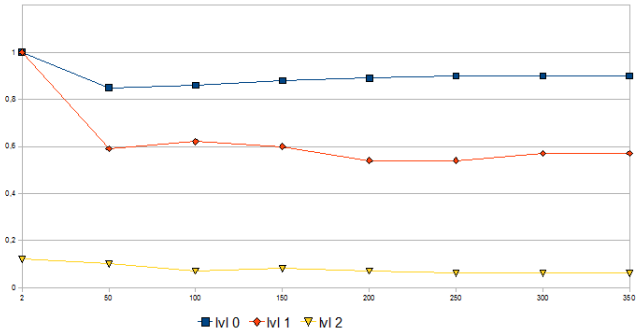


Fig. 11. Inference accuracy on different granularity levels

Granularity Levels As important aspect of the given representation and the ability to perform inference on different levels of granularity, we were able to test the inferences on the given three levels by their accuracy. Therefore we see in figure 11 an expected behavior if the granularity level is more abstract less errors occur. On the finest granularity level the average error is relatively high with 89% but this value does not consider that maybe not the right hypothesis has been chosen but a very similar hypothesis. This assumption can be confirmed by looking at the more abstract granularity levels. On the next abstraction level the average error is 59% selecting between 30 hypotheses and on the most abstract level the error could be reduced to an average of 7%.
This behavior offers the opportunity to specify a minimal certainty for the prediction and to perform inferences of a dynamic granularity level. By adopting the level of granularity a hypothesis can be searched exceed the given certainty. So instead of defining a fixed granularity level a minimum certainty is used to automatically determine an ideal granularity level.

Over-Generalisation A seldom and not preferable property of the RHMM is the over-generalisation in some cases. If only a very small amount of reference data is available for one state but a very large amount of data is available for another but very similar state, the state with the few reference data will be mostly neglected during inference. This is not always preferable, because the few data could be a more appropriate basis for the inference of this state. During the tests in the *RoboCup* domain this non-preferable mechanism was nearly neglectable cause the reference data was relatively good distributed.

Complexity The complexity of the presented inference can be reduced to the complexity of a *HMM* by precalculating the model parameters considering the relational dependencies. The precalculation effort on the other hand is highly dependent to the complexity of the model structure, especially the amount of attributes for a relation and the attributes' domain complexity. To indicate how complex the precalculations and the inference it-self are, the following tables show the used time on a *Athlon 2400 XP-M*:

Test	Time Ø	Tests	σ	Min	Max
Training	56,45	1000	0,16	50	70
LookUpTable	3,68	1000	0,15	0	10
Inference	272,82	14900	0,04	260	430

Fig. 12. *RHMM-RANK* inference without precalculations

Test	Time Ø	Tests	σ	Min	Max
Training	68,01	1000	0,15	50	80
LookUpTable	704,23	1000	0,24	680	730
Inference	2,99	14900	0,03	0	10

Fig. 13. *RHMM-RANK* inference with precalculations

Fig. 12 shows the time used for the inference without any precalculations with two evidences. Fig. 13 shows the same measurements with precalculations. The tables show that the precalculations for the given model structure can be done within a very short period of time ($704,23ms$) and decrease the used inference time significantly (from $272,8ms$ to $2,99ms$).

4 Discussion and Future Work

One of the most demanding task to be accomplished in scene interpretation based on physically grounded robots is the handling of uncertainty. Uncertainty arises essentially from two different sources: (1) perception: sensor noise and (2) ambiguity and missing (and even false) information in the process of interpretation. In this paper we presented approaches to both classes of problems. In the first part (section 2) we described an approach to localization[7] based on qualitative ordering information that does not rely on probabilistic inference. In addition we showed that a strictly propositional representation can be obtained by abstraction into qualitative (predefined) frames of reference. Nevertheless, even if the use of probabilistic inference (and representation) can be avoided in terms of localization and the generation spatial world model, the problem of uncertainty has to be addressed at least due to the inherent ambiguity of the interpretation process and the missing information required for monotonic logical inference. Therefore, in the second part we presented the *relational hidden Markov model* (RHMM) which is based on the well-established *HMMs* and the *RMM*, and we showed how it could be applied to *spatio-temporal* reasoning. While prediction is an important inference for dynamic scene interpretation it is strongly embedded into the overall interpretation (reasoning) process. Depending on the specific interpretation and the information available the requirements will differ significantly with respect to the required precision and the validity of the generated prediction hypotheses. We showed that the *RHMM* can be used efficiently to address this problem. In the RoboCup-domain the use of *RHMM* leads to an increased inference accuracy with a minimally increased calculation effort. In this domain the *RHMM* could predominantly outperform the well-known *HMM* in inference accuracy, but it was also shown that the required inference mechanism is more complex than the one of HMM. However, by pre-calculating the models' relational parameters the inference complexity could (in the RoboCup-domain) practically be reduced to the complexity of a *HMM*. This makes the model especially interesting for time critical

[7] The presented approach addresses the *global localization*-problem and is not limited to position tracking (in terms of Thrune's classification [17]).

domains with *spatio-temporal* representations based on noisy sensor information. The increased inference accuracy results from the ability of modeling domain specific information like similarities between discrete states.

To further improve the inference accuracy of the *RHMM* the attribute domains can probably be determined automatically to represent the underlying environment domain in an optimal way. Perhaps even the states themselves could be determined in such a way. Furthermore, the application in different domains could be interesting in order to gather more results for different types of domain representations.

References

1. J. F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, pages 123–154, 1984.
2. C. Anderson, P. Domingos, and D. Weld. Relational Markov Models and their Application to Adaptive Web Navigation, 2002.
3. E. Bourque and G. Dudek. Automated imagebased mapping, 1998.
4. Eliseo Clementini, Paolino Di Felice, and Daniel Hernandez. Qualitative representation of positional information. *Artificial Intelligence*, 95(2):317–356, 1997.
5. H. Durrant-Whyte, S. Majumder, S. Thrun, M. de Battista, and S. Schelling. A bayesian algorithm for simulaneous localization and map building. In *Proceedings of the 10'th International Symposium on Robotics Research (ISRR'01))*, Lorne, Australia, 2001. AAAI Press/MIT Press.
6. C. Freksa. Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54(1):199–227, 1992.
7. Fikes R. E. Garvey T. D. Nilsson N. J. Nitzan D Tenenbaum J. M. Hart, P. E. and B. M. Wilber. Artificial intelligence - research and applications. Technical report, Stanford Research Institute, December 1972.
8. T.S. Levitt and D.T. Lawton. Qualitative navigation. *AI*, 44(3):305–361, August 1990.
9. A. Miene. *Raeumlich-zeitliche Analyse von dynamischen Szenen*. PhD thesis, Universitaet Bremen, 2004.
10. A. Miene, A. Lattner, U. Visser, and O. Herzog. Dynamic-preserving qualitative motion description for intelligent vehicles, 2004.
11. A. Miene, U. Visser, and O. Herzog. Recognition and prediction of motion situations based on a qualitative motion description. In Polani et al., editor, *RoboCup 2003: Robot Soccer World Cup VII*, volume 3020 of *LNCS*, pages 77–88. Springer, 2004.
12. A. Miene and T. Wagner. Static and Dynamic Qualitative Spatial Knowledge Representation for Physical Domains. Vol. 2/06:pp. 109–116, 2006.
13. C. Schlieder. Representing visible locations for qualitative navigation. pages 523–532. In: Qualitative reasoning and decision technologies, N. Piera-Carrete & M. Singh (Eds.) CIMNE Barcelona [http://www.iig.uni-freiburg.de/cognition/members/cs/cs-pub.html], 1993.
14. C. Schlieder. Ordering information and symbolic projection, 1996.
15. Christoph Schlieder. *(Doctoral Thesis) Anordnung und Sichtbarkeit - Eine Charakterisierung unvollstndigen rumlichen Wissens*. PhD thesis, University Hamburg, Department for Computer Science, 1991.
16. S. Thrun. Robotic mapping: A survey. In G. Lakemeyer and B. Nebel, editors, *Exploring Artificial Intelligence in the New Millenium*. Morgan Kaufmann, 2002. to appear.
17. S. Thrun, D. Fox, W. Burgard, and F. Dellaert. Robust monte carlo localization for mobile robots. *Artificial Intelligence*, 128(1-2):99–141, 2000.
18. S. Thrun, D. Fox, W. Burgard, and F. Dellaert. Robust monte carlo localization for mobile robots. *Artificial Intelligence*, 128:99–141, 2001.
19. T. Wagner and K. Huebner. An egocentric qualitative spatial knowledge representation based on ordering information for physical robot navigation. In *Proceedings of Paper the RoboCup VII Symposium*, 2004.
20. Thomas Wagner. *(Doctoral Thesis) Qualitative sicht-basierte Navigation in unstrukturierten Umgebungen (submitted)*. PhD thesis, University Bremen, FB-3, TZI, AG-KI, 2005.