

Bayesian Reasoning on Qualitative Descriptions from Images and Speech

Gudrun Socher¹

Gerhard Sagerer²

Pietro Perona³

¹Xerox PARC, 3333 Coyote Hill Rd., Palo Alto, CA 94304, USA

²Universität Bielefeld, Technische Fakultät, Postfach 100131, 33501 Bielefeld, Germany

³California Institute of Technology, 136-93, Pasadena, CA 91125, USA

e-mail: gudrun@techfak.uni-bielefeld.de

Abstract

Image understanding denotes not only the ability to extract specific, non-numerical information from images, but it implies also reasoning about the extracted information. We propose a qualitative representation for image understanding results which is suitable for reasoning with Bayesian networks. Our representation is not purely qualitative but enhanced with probabilistic information to represent uncertainties and errors in the understanding of noisy sensory data. The probabilistic information is then supplied to a Bayesian networks in order to find the most plausible interpretation.

We apply this approach for the integration of image and speech understanding to find objects in a visually observed scene which are verbally described by a human. Results demonstrate the performance of our approach.

1 Introduction

The representation of image understanding results is an important issue. With image understanding, we mean the extraction of symbolic or qualitative, non-numerical information from images and the reasoning about the extracted information. Understanding results are often affected by errors or ambiguities due to noisy data or erroneous intermediate results. We propose therefore a qualitative representation for image understanding results which is enhanced by probabilistic information characterizing the reliability of the results. A Bayesian network approach is then applied using this representation to find the most plausible result.

We apply this technique to extract relevant qualitative information from images and speech for the integration of image and speech understanding in a system for natural human-computer interaction.

Our scenario is the cooperative assembly of toys using the wooden toy construction kit *Baufix* (see Fig. 3 for objects in our scenario). A human plays the role of an instructor and gives verbal instructions to the system. The system is equipped with a microphone as well as a stereo camera to observe the scene. Using the information from

both, speech and images, the system should understand the given instructions, relate them to the objects in the scene, and carry out the instructions.

Necessary components in the system are

- speech understanding,
- image understanding,
- an inference machine in order to integrate image and speech understanding for the identification of the *intended object*¹ and the command execution.

We integrate image and speech understanding on a symbolic level. Hence, we extract our probabilistically enhanced *qualitative descriptions* from images and speech and reason upon them. The qualitative descriptions characterize objects in terms of their type, color, size, and shape as well as spatial relations relative to other objects. Fig. 1 sketches the components and their interaction in this system.

In this paper, we focus on the extraction of qualitative descriptions from images and explain our inference mechanism using Bayesian networks. The following subsection refers to related work. In Section 2, we explain our representation of *qualitative descriptions*, and Section 3 outlines the computation of qualitative descriptions from images. The Bayesian inference mechanism is described in Section 4. Important aspects are here not only the network architecture but also the estimation of the conditional probability tables. Results and a discussion of our approach conclude this paper.

1.1 Related Work

Our work was inspired by a number of approaches investigating the integration of computer vision and natural language (Wahlster, 1989; Mc Kevitt, 1994; Srihari, 1994) as well as approaches generating conceptual descriptions

¹We call an object which is referred in an instruction the *intended object*.

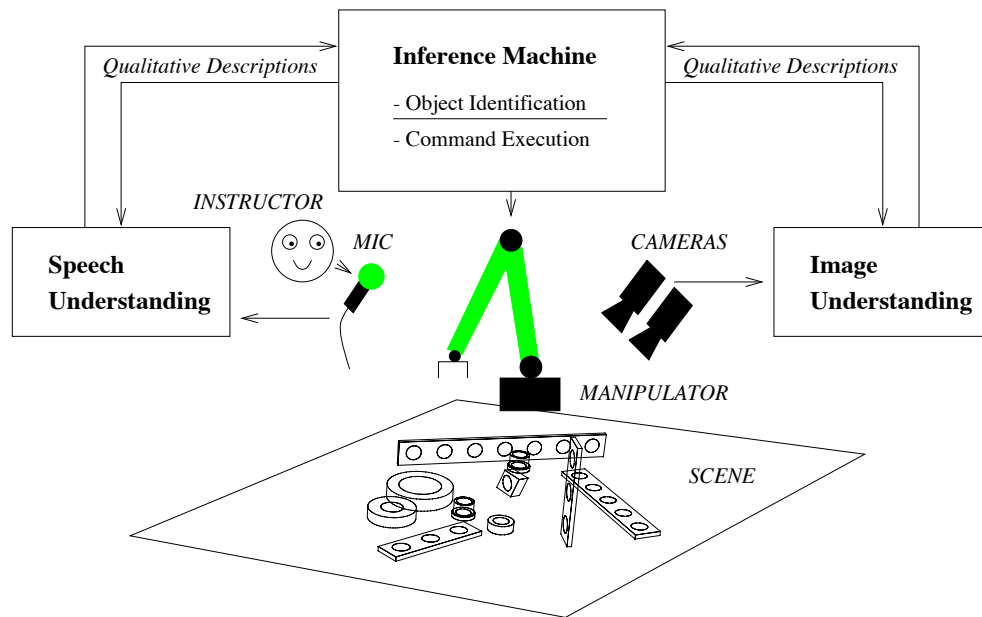


Fig. 1: Overview of our image and speech understanding system: The image and speech understanding modules derive qualitative descriptions from speech and visual input data. A Bayesian inference machine is used for for object identification and command execution.

from images (Nagel, 1988; André et al., 1988; Toal & Buxton, 1992; Kollnig & Nagel, 1993) and systems which visually observe the scene to enable natural human-computer interaction (Tsotsos et al., 1997).

The group around Wahlster (1989) and André et al. (1988) investigates the integration of vision and natural language in various systems. An example is the system SOCCER (e.g André et al., 1988) which simultaneously analyses and describes short scenes of soccer games like live radio reports. This involves perceiving and interpreting the locations and movements of the ball and players in order to select how to describe the game and in what sequence. ‘Probability clouds’ and ‘typicality fields’ are used to probabilistically describe the image understanding results.

The PICTION system (Srihari, 1994) uses captions to identify human faces in an accompanying photograph in order to explore the interaction of textual and visual information. A face locator is used to segment face candidates from an image at different resolutions of the original image and the edge image. Constraints for the face recognition are then generated from the semantic processing of the caption. Picture-specific information is extracted to generate contextual (e.g., the name), characteristic (e.g., the gender), and locative or spatial identifying constraints. The constraints guide the processing of the picture to provide a semantic interpretation.

XTRACK (Nagel, 1988; Koller et al., 1993; Kollnig & Nagel, 1993) is an example of an image understanding system which performs fully automatically all necessary steps from low-level image analysis to conceptual descriptions of moving vehicles in traffic scenes. Trajectories of moving vehicles are extracted from image sequences and are conceptually described by motion verbs. Fuzzy sets are used to represent the connections between trajectory attributes and motion verbs. The admissible sequences of activities of an agent are modeled by hierarchical ‘situation graphs’.

The goal of the PLAYBOT project (Tsotsos et al., 1997) is to provide a directable robot which may enable physically disabled children to access and manipulate toys. The robot possesses a robotic arm and hand, a stereo color vision robot head, and a communication panel. The child gives commands on the communication panel (execute actions with toys). The system is able to visually explore the room to perceive objects and their location and to execute the given commands. The scenarios of PLAYBOT and our system are similar. However, PLAYBOT is focused on the visual perception of the environment. The command language is rather simple and therefore less attention is drawn on the representation of understanding results and on inference processes in order to identify the intended object or action.

2 Representing Qualitative Descriptions

Harnad (1987) declares categories as the basic representational units, and all qualitative entities result from discretizations of the continuous signal space in categories. Following Harnad's definition, Medin & Barsalou (1987) distinguish between two different kinds of categories: (1) all-or-none categories and (2) graded ones. There are two subtypes of all-or-none categories: (1a) In 'well-defined' categories, all members share a common set of features and a corresponding rule defines them as necessary and sufficient conditions for membership. (1b) In 'defined' (but not well-defined) categories the features do not need to be shared by all members, and the rule can be an either/or one. Graded categories (2) are not defined by an all-or-none rule at all, and membership is a matter of degree.

The boundaries between categories should be placed in such a way that there are qualitative resemblances within each category and qualitative differences between them. But nevertheless, a qualitative property is always more or less true, or applicable, to a physical object or a set of objects. The boundaries between categories may show fuzziness. Therefore, we follow Medin & Barsalou (1987) and use graded categories as basic representational unit. The degree of membership is a fuzzy or probabilistic value which represents a goodness of fit between a category and the underlying numerical data. However, categories are not necessarily disjoint; they may overlap, for example, an object can be colored bluish green. Thus, we characterize each *property* by a finite number of graded *categories* (e.g., white, red, yellow, orange, blue, etc.). This leads us to the definition of *qualitative descriptions* as representational scheme:

A **qualitative description** characterizes an entity by a set of *properties*. Each property is described as a vector of graded *categories*.

To describe properties more formally, each qualitative property is a function Q_u or Q_b :

$$Q_u(\mathbf{p}, \mathbf{o}) = \mathbf{u} \quad \text{or} \quad Q_b(\mathbf{p}, \text{IO}, \text{RO}) = \mathbf{u}, \quad (1)$$

for unary relations (Q_u) and for binary relations (Q_b) such as spatial relations. \mathbf{o} denotes the object involved in an unary relation and IO and RO stand for the *intended object* and the *reference object* in spatial relations (see Subsection 3.3). \mathbf{u} is a vector which represents the fuzzy degrees of membership that are assigned to each category of a property space. u_i characterizes how well category i of property \mathbf{p} fits for the given object \mathbf{o} or object pair (IO, RO).

Here are two examples:

1. The property *color* contains the following categories:

$$P_{color} = (\text{red, yellow, orange, blue, green, purple, wooden, white})^T.$$

Then \mathbf{u} is a vector of dimension 8 and

$$Q_u(\text{color}, \text{rhomb-nut}) = (0.4, 0.3, 0.8, 0.1, 0.09, 0.2, 0.15, 0.05)^T.$$

This means that the object *rhomb-nut* is most likely to be orange (cf. Fig. 2). The color orange is also somewhat red as well as dark yellow, and thus the degrees of membership for the categories red and yellow are higher than for the other color categories.

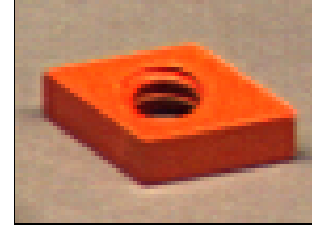


Fig. 2: A rhomb-nut.

2. The property *spatial relation* consists of the projective relations left, right, above, below, behind, and in-front, and thus

$$P_{spatial\ relation} = (\text{left, right, above, below, behind, in-front})^T.$$

Then is, for example,

$$Q_b(\text{spatial relation}, \text{rhomb-nut}, \text{socket}) = (0.18, 0.04, 0.03, 0.05, 0.74, 0)^T.$$

This example is taken from Table 5 (IO = 1, RO = 4). It describes that the rhomb-nut is behind and slightly left of a socket.

This representation has the following benefits:

- The uncertainty from recognition or detection process can easily be represented.
- Overlapping meanings and concurring hypotheses can be represented.
- It captures the degree of membership for each category of a property space. No irreversible, strict decisions have to be made to compute the qualitative value of a property in an early stage of an understanding process.

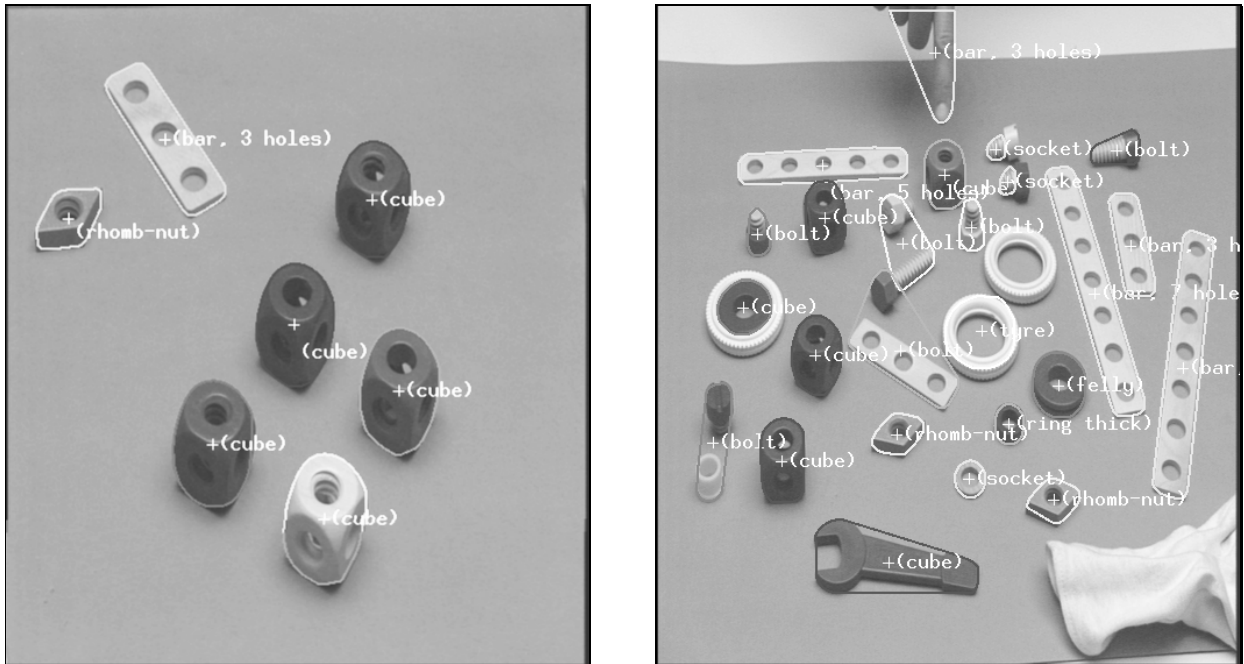


Fig. 3: Object recognition results from two scenes: All objects of the simple scene (left image) are recognized correctly, whereas the object recognition module has more difficulties with the complex scene (right image).

- A decision, which single category is chosen in a property space, can be postponed to a later stage and taken using a suitable decision calculus. Furthermore, other information that might be available later, e.g., about other properties or the scene context, can be taken into account in the process of making the decision for the final qualitative system output.
- Object specifications can be of a wide variety. Using a decision calculus, the understanding of verbal object specifications is rather flexible in reacting to the choice of properties which are specified in an instruction. Even partially false specifications can be handled.

3 Computation of Qualitative Descriptions

We compute qualitative descriptions of objects from images in terms of ‘type’, ‘color’, and ‘spatial relations’ relative to other objects.

3.1 Object Recognition

The most obvious description of an object is naming its type or object class. In the Bafix domain, there are objects of 20 different object types. Object recognition is carried out by a hybrid approach combining neural and semantic networks (cf. Heidemann et al., 1996). The neural network generates object hypotheses, which are either verified or rejected by a semantic network approach (see Socher, 1997 for details).

Two examples of object recognition results are shown in Fig. 3. Here the best scored results are displayed. The qualitative description of the object type captures the scored recognition results.

3.2 Color

Color is a dominant feature for object descriptions. Subjects prefer to specify the visually most salient feature (Herrmann & Deutsch, 1976), and this is often the color of an object. In the Bafix scenario, the objects are colored with bright and clearly distinct elementary colors.

We use a rather simple color classification approach. A pixelwise color classification is performed by a polynomial classifier of 6th degree on HSI color images. Subsequent smoothing operations and region labeling lead to color segmented images. Currently, the lighting conditions are fixed, and the limited set of Bafix colors is pretty well distinguishable.

The computed color classifications for the image region of an object are assigned to the qualitative color description of that object. So far, no classification score is recorded. Therefore, the qualitative color description is initialized with the score 1 for the category of the classified color and with the score 0.01 for all other color categories. The fuzzy vector is then normalized.

3.3 Spatial Relations

Describing the spatial location of an object is another means for specifying an object. Our computational model

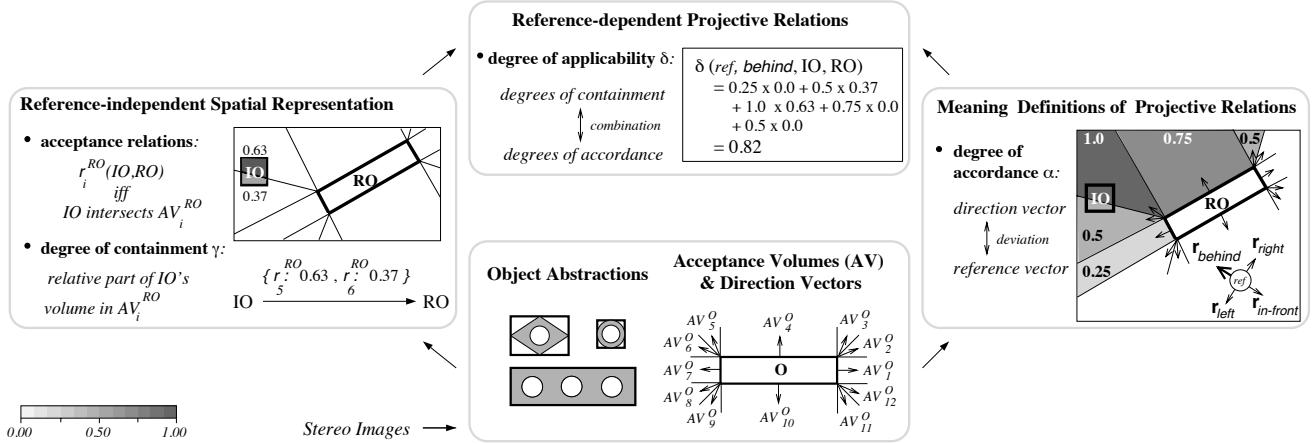


Fig. 4: Computation of scored projective relations for object pairs: The principles of the computation are demonstrated for 2D objects in 2D space. An object-specific partitioning of 2D space into 12 acceptance volumes is chosen. As an example the generation of the relation *behind* is shown for the objects IO and RO w.r.t. the reference frame *ref*.

for spatial relations is designed to compute the binary projective relations *left*, *right*, *above*, *below*, *behind*, and *in-front* for the *intended object* (IO) relative to the *reference object* (RO) in 3D given a *reference frame* (Fuhr et al., 1997). Input data to the computational model is a 3D reconstruction of the scene which is computed from stereo images based on the results from object recognition (see Socher, 1997).

Objects in 3D are abstracted by bounding boxes which are collinear to the object’s principal axes. A finite number of acceptance volumes AV_i^O is associated with each object O . These are infinite open polyhedra bound to the sides, edges, and corners of the object partitioning the 3D space. A direction vector $d(AV_i^O)$ corresponds to each acceptance volume. It roughly models the direction to which an acceptance volume extends in space. The object-specific partitioning is motivated by the assumption that the object itself may influence the way the surrounding space is perceived independently of specific reference frames.

The computation of relations from objects is a two-layered process. In the first layer, a *reference-independent* spatial representation is computed. Each acceptance volume induces a binary *acceptance relation* r_i^O that expresses whether an object P intersects with AV_i^O . Acceptance volumes are scored by calculating the corresponding *degree of containment*:

$$\gamma(P, r_i^O) = \frac{\text{vol}(P \cap AV_i^O)}{\text{vol}(P)}. \quad (2)$$

Thus, the relation between two objects P and O can be *reference-independently* expressed by a set of scored acceptance relation symbols with non-zero degree.

Furthermore, *reference-dependent meaning definitions*

of relations rel^2 w.r.t. certain ROs and a given reference frame $\text{ref} = \{\mathbf{r}_{\text{left}}, \mathbf{r}_{\text{right}}, \dots, \mathbf{r}_{\text{in-front}}\}$ are also calculated in the first layer. The meaning definition $\text{def}(\text{ref}, \text{rel}, \text{RO})$ is given as the set of the symbols of all acceptance relations r_i^{RO} whose direction vector differs less than 90° from the corresponding reference vector \mathbf{r}_{rel} . The membership of an acceptance relation (symbol) to a meaning definition is scored by its *degree of accordance*:

$$\alpha(\text{ref}, \text{rel}, r_i^{\text{RO}}) = 1 - 2 \cdot \frac{\arccos(d(AV_i^{\text{RO}}) \cdot \mathbf{r}_{\text{rel}})}{\pi}. \quad (3)$$

These two scored symbolic reference-independent and reference-dependent descriptions are the basis for the computation of reference-dependent relational expressions for IO-RO pairs in the second layer. The basic idea is, that the relation *rel* is applicable for an IO-RO pair w.r.t. a reference frame *ref* if at least one of the acceptance relations in $\text{def}(\text{ref}, \text{rel}, \text{RO})$ holds between IO and RO. The *degree of applicability* $\delta(\text{ref}, \text{rel}, \text{IO}, \text{RO})$ of *rel* varies gradually:

$$\delta(\text{ref}, \text{rel}, \text{IO}, \text{RO}) = \sum_{\substack{r_i^{\text{RO}} \in \\ \text{def}(\text{ref}, \text{rel}, \text{RO})}} \alpha(\text{ref}, \text{rel}, r_i^{\text{RO}}) \cdot \gamma(\text{IO}, r_i^{\text{RO}}). \quad (4)$$

Fig. 4 illustrates the steps of this computation. For easier visualization the steps are shown for 2D objects in 2D space.

The table in Fig. 5 shows computed degrees of applicability of spatial relations for objects in the depicted scene. The reference frame is assumed to correspond with the cameras’ view of the scene. The table demonstrates that the results are very promising keeping in mind that they have

²*rel* is a generic representative of the set of spatial relations

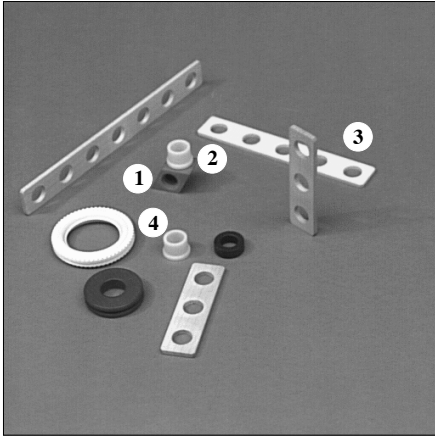


Fig. 5: Example of computed spatial relations for the numbered objects in the depicted scene: The maximum applicability degree per RO-IO pair is highlighted in bold. The chosen reference frame takes the position of the camera to allow for an easy verification of the results from the image of the scene.

IO	RO	left	right	above	below	behind	in-front
2	1	0.17	0.10	0.51	0	0.15	0.09
1	2	0.12	0.01	0	0.86	0.04	0.01
3	1	0	0.60	0.04	0.01	0.39	0
1	3	0.26	0	0.13	0.07	0	0.64
4	1	0	0.20	0.04	0.07	0	0.75
1	4	0.18	0.04	0.03	0.05	0.74	0
3	2	0	0.52	0	0.23	0.32	0
2	3	0.24	0	0.23	0	0	0.65
4	2	0.02	0.06	0	0.17	0	0.79
2	4	0.09	0.03	0.17	0	0.78	0
3	4	0	0.34	0	0	0.65	0
4	3	0.24	0	0.19	0	0	0.66

been computed from slightly erroneous 3D object poses reconstructed from *real* stereo images.

4 Reasoning using Bayesian networks

The computation of qualitative descriptions forms a transition between numerical input data and symbolic, abstract descriptions. To actually reason upon the qualitative descriptions and to identify objects referred to by instructions, we use Bayesian networks. Bayesian networks are designed to reason under uncertainty. In our application errors in the recognition process, ambiguities, slips of the tongue, insufficient specifications, and other inaccuracies, may corrupt the qualitative descriptions.

The identification of the object(s) in the visible part of the scene, which are referred to in an instruction, is important in our image and speech understanding system. However, our goal is the integration of image and speech understanding as well as the inference of constraints for missing information in either type of qualitative descriptions and the generation of object descriptions as feed-back to the human instruction. Bayesian networks are very well suited for this task. They offer the possibility of bottom-up, top-down, and mixed mode reasoning under uncertainty. This allows us to infer constraints and to deal with data in real and noisy environments.

In this section, we first describe Bayesian networks and the design of our Bayesian network for object identification. Crucial for Bayesian networks is the modeling of the conditional probabilities. We estimate them based on prior domain knowledge (see Subsection 4.2.3). We carried out a questionnaire on the World Wide Web to investigate the use of size and shape properties in our domain which is reported in Subsection 4.2.4.

4.1 Bayesian Networks

Bayesian networks are directed acyclic graphs in which nodes represent random variables and arcs signify the existence of direct causal influences between the linked variables (Pearl, 1988). Bayesian networks are an explicit representation of the joint probability distribution of a problem domain (a set of random variables X_1, \dots, X_n), and they provide a topological description of the causal relationships among variables. If an arc $X_i \rightarrow X_j$ is established then the probability of each state of X_i depends on the state distribution of X_j .

A conditional probability table (CPT) is associated with each arc. It provides conditional probabilities of a node's possible states given each possible state of the parent which is linked by this arc. The CPTs express therefore the strength of the causal influences between the linked variables. If a node has no parents, the prior probabilities for each state of the node are given in the CPT.

Three parameters are attached to each node representing the belief in the state, as well as diagnostic (λ) and causal (π) support from incoming and outgoing links, respectively. Beliefs are updated taking into account both parents and children. We employ the propagation algorithm for trees suggested by (Russell & Norvig, 1995).

4.2 Object Identification using Bayesian Nets

The design of the Bayesian network for object identification was guided by the following requirements:

- Decisions should account for uncertainty in the data as well as uncertainties in the recognition and interpretation processes.
- Data and results from psycholinguistic experiments should be integrated easily.

- It should be possible to model lacks of performance of the recognition modules.
- The system should be able to identify objects from unspecific and even partially false instructions/object descriptions.
- It should be possible to infer constraints for missing information of any type.

These requirements are satisfied best, when we determine from all detected objects of a scene the one(s) which are most likely referred to in an instruction. This means that the object(s) with the highest joint probability of being part of the scene, and being referred to, are the intended object(s). In this way, we can incorporate the uncertainties which are involved in the recognition as well as the understanding processes. We can also account for erroneous specifications as we do not require a perfect match of observed and uttered features but only the highest probability of a match of these features among all detected objects. Furthermore, we identify an object in the context of the scene and of all uttered features, and we do not just pairwise compare uttered and observed features.

We designed our Bayesian network (cf. Fig. 6) according to these guidelines. It is a tree-structured network. The root node **identified object** represents the intended object. The dimension of this node is 23 as there are 23 different objects in our domain. Thus, we estimate for each object of the domain the probability or likelihood of being intended. The children of the root node represent the **scene** and the **instruction**. The dimension of these nodes is again 23 and they represent for each object either the probability of being part of the current scene or of being named in the instruction.

The children of the node **instruction** are instantiated with the qualitative descriptions resulting from speech understanding. If a property is not specified in an instruction, then the diagnostic support of the corresponding node is set to $1_{n \times 1}$, where n is the dimension of the property. This means that the vector contains a 1 in every component.

The observed objects are represented by the nodes **object₁, ..., object_m**. Each node has again dimension 23 and represents for an observed object all possible object categories and their likelihood of characterizing it. Typically, one component has a high belief, and the belief for all other components is low. The qualitative type and color descriptions are used to instantiate the nodes **type** and **color** for each observed object.

4.2.1 Object Identification

The intended object(s) are identified in the following way: First, after the initialization of the network, only $\lambda(\mathbf{scene})$

and $\pi(\mathbf{io})$ are propagated through the network. ($\pi(\mathbf{io})$ is the vector of prior probabilities for the node **identified object**. It is set to $1/23$ in every component.) The resulting beliefs for the objects nodes, prior to normalization, characterize the certainty of detection in the context of the scene. We call this value **offset** and compute it as

$$\mathbf{offset}_j = \pi_{init}(\mathbf{object}_j) \cdot \lambda_{init}(\mathbf{object}_j). \quad (5)$$

Incoming evidence is then propagated bottom-up and top-down through the network. The belief of the scene node results from incoming evidence as well as from top-down propagated messages from the node **identified object**. Hence, the belief of the scene node represents for each domain object the joint probability of being part of the scene and being intended. Messages from the scene node are propagated top-down to the nodes **object₁, ..., object_m**.

After the propagation of evidence obtained from image and speech understanding, the beliefs of the nodes **object₁, ..., object_m** are again taken prior to normalization. We define for each object j the possibility τ_j of being intended as

$$\tau_j = \pi(\mathbf{object}_j) \cdot \lambda(\mathbf{object}_j). \quad (6)$$

For each object, a likelihood value η is taken

$$\eta_j = \max_i((\tau_i)_j) - \max_i((\mathbf{offset}_i)_j). \quad (7)$$

η_j is the difference of the maximal component of τ_j and the maximal component of \mathbf{offset}_j . This gives us one likelihood value for each observed object. The identified object should be the most likely intended one. Therefore, we use a little statistical analysis to find this/these object(s). We compute the mean μ and the standard deviation σ of the set of likelihood values η_j . Objects with likelihood values below $\mu - \sigma$ are excluded from the statistical analysis. We define the selection criterion as

$$\text{object } j \text{ identified if } \begin{cases} \sigma < \text{threshold and } \eta_j > 0 \\ \sigma \geq \text{threshold and } \eta_j > \mu + \sigma. \end{cases} \quad (8)$$

This gives us all outliers with a level of confidence of 0.69. Thus, we get all those objects which have a significantly higher belief of being intended than the other objects. A level of confidence of 0.69 is not too low for our scenario. We have to cope with noisy and erroneous data. And we want the system to identify an object rather than to report “no objects found” even if a lot of uncertainty is involved. The instructor can correct false identifications, and it is more convenient to correct once in a while a false identification than to repeat instructions multiple times until the system has finally found the intended object.

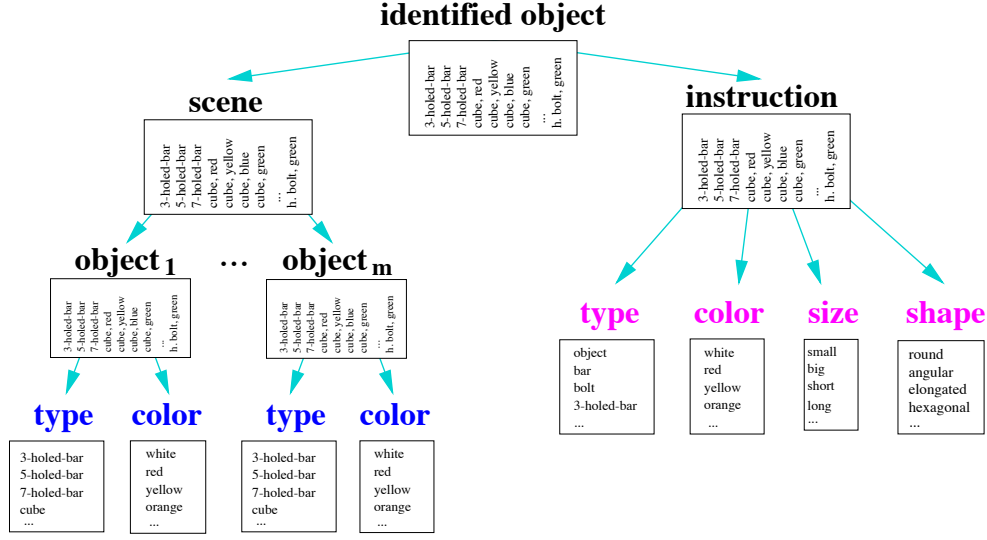


Fig. 6: Structure of the Bayesian network for object identification.

4.2.2 Spatial Relations

So far, we have only described how we identify objects from the unary properties type, color, size, and shape. Binary, spatial relations are another means for referring to objects. Especially when there are multiple objects of the same type in the scene, then it might be the only way to uniquely specify an object. But a tree-structured Bayesian network is not well suited to handle binary relations. Furthermore, we envision to model the spatial configuration of a scene as relation graph such as proposed by Fuhr et al. (1993). We think that a graph structure is the best way to handle the complexity of binary or even higher dimensional relations.

We propose therefore at the current state of the system implementation a very simple method for handling object identifications that use spatial relations. $\mathbf{spat_rel}_{inst}$ is the qualitative description of the uttered spatial relation(s). The following steps are executed:

1. Identify all candidates for possible intended objects (IO) based on type, color, size, and shape, if named.
2. Identify all candidates for possible reference objects (RO) based on type, color, size, and shape, if named.
3. Compute spatial relations for all IO/RO candidate pairs as explained in Section 3.3.

- (a) Compute for each IO/RO candidate pair the Euclidean distance of the vector of computed spatial relations $\delta(ref, IO, RO)$ and $\mathbf{spat_rel}_{inst}$.

$$r_d = \|\delta(ref, IO, RO) - \mathbf{spat_rel}_{inst}\|_2 \quad (9)$$

- (b) Compute $s = \eta_{IO} * \eta_{RO} / r_d$.

4. The IO/RO candidate pair with the greatest s is identified.

4.2.3 Use of prior Knowledge

The conditional probability tables are estimated from results of psycholinguistic experiments and error analyses of the understanding modules.

In a first psycholinguistic experiment 10 subjects named objects verbally. Images of scenes with Bauxif objects were presented to the subjects on a computer screen. In each image, one object was marked with a pink arrow and the subjects were told to name this object using an instruction in the form of an instruction such as “give me *the object*” or “take *the object*”. From this experiment, 453 verbal object descriptions were collected. The properties named the most in these instructions are the type and the color of the objects. We estimated from this data

- $P(\mathbf{type}_{inst.[j]} | \mathbf{instruction}[i]) = \frac{\# \mathbf{type}_i \text{ was named when object}_i \text{ was shown}}{\# \text{ object}_i \text{ was shown}}$,
- $P(\mathbf{color}_{inst.[j]} | \mathbf{instruction}[i]) = \frac{\# \mathbf{color}_j \text{ was named when object}_i \text{ was shown}}{\# \text{ object}_i \text{ was shown}}$.

We denote the i^{th} component of a vector \mathbf{x} as $\mathbf{x}[i]$. We want to avoid conflicting indices with this notation known from programming languages.

The second experiment was a questionnaire in the World Wide Web for size and shape of the objects (see Sec-

tion 4.2.4). It is used for the following conditional probabilities:

- $P(\mathbf{size}_{inst.[j]}|\mathbf{instruction}[i]) = \frac{\#size_j \text{ was named when object}_i \text{ was shown}}{\# \text{ object}_i \text{ was shown}}$,
- $P(\mathbf{shape}_{inst.[j]}|\mathbf{instruction}[i]) = \frac{\#shape_j \text{ was named when object}_i \text{ was shown}}{\# \text{ object}_i \text{ was shown}}$.

The performance of the speech understanding system was not evaluated here. Therefore, confusions which may occur in speech understanding are not yet modeled.

Another series of experiments tests the image understanding modules. Object recognition was performed on 11 images of different scenes which were taken under constant lighting conditions but with different focal lengths. The conditional probabilities $P(\mathbf{type}_{object}|\mathbf{object}_k)$ are estimated as

- $P(\mathbf{type}_{object}[j]|\mathbf{object}_k[i]) = \frac{\#type_j \text{ was detected when object}_i \text{ was depicted}}{\# \text{ object}_i \text{ depicted}}$,

k denotes the k^{th} object in the scene. The conditional probability tables are the same for all objects.

The color classification performance was also evaluated. This gives us a conditional probability table

$$P(\mathbf{color \text{ classif.} | \text{pixel color}}) = [P(\mathbf{color}_j \text{ classif.} | \text{pixel color}_i)]_{i,j},$$

with $P(\mathbf{color}_j \text{ classif.} | \text{pixel color}_i) = (\# \text{ color}_j \text{ classified} / \# \text{ pixel with color}_i)$. But we need the conditional probabilities $P(\mathbf{color}_{object}[j]|\mathbf{object}_k[i])$. We estimate these by using an ideal transition matrix $P(\mathbf{color}|\mathbf{object})$ and multiplying it with $P(\mathbf{color \text{ classif.} | \text{pixel color}})$. We set $P(\mathbf{color}_j|\mathbf{object}_i) = \alpha$ if the color of object _{i} is color _{j} and ε otherwise, where α is a probability near 1 but normalized so that $P(\mathbf{color}|\mathbf{object}_i) = 1$, and ε is a very small probability. Therefore,

- $P(\mathbf{color}_{object}[j]|\mathbf{object}_k[i]) = [P(\mathbf{color}|\mathbf{object}) \cdot P(\mathbf{color \text{ classif.} | \text{pixel color}})]_{i,j}$.

The transitions between **identified object** and **scene** and **instruction** as well as between **scene** and all objects **object _{k}** are considered as unbiased and so the conditional probabilities were set to unit matrices where the zero entries are replaced by very small probabilities:

- $P(\mathbf{scene}[j]|\mathbf{identified \ object}[i]) = P(\mathbf{instruction}[j]|\mathbf{identified \ object}[i]) = \begin{cases} \alpha & \text{if } i = j \\ \varepsilon & \text{otherwise,} \end{cases}$
- $P(\mathbf{object}_k[j]|\mathbf{scene}[i]) = \begin{cases} \alpha & \text{if } i = j \\ \varepsilon & \text{otherwise.} \end{cases}$

4.2.4 Questionnaire about Size and Shape in the World Wide Web

So far, we described how to extract the qualitative properties, type, color, and spatial relations from numerical data. But what about the size and shape of objects? Size and shape are not named in the instructions as often as the color but nevertheless, our system should be able to cope with size and shape specifications. This leads to the question whether there are any classification schemes for size and shape. Are there functions of the metric size and shape of the objects (e.g., volume, diameter)? Or do other mechanisms apply? We were unable to answer these questions and decided therefore to start a questionnaire in the World Wide Web (WWW) in order to collect empirical data about which size and shape categories subjects associate with the objects in our domain. The World Wide Web is a means to reach many people and to acquire large sets of data.

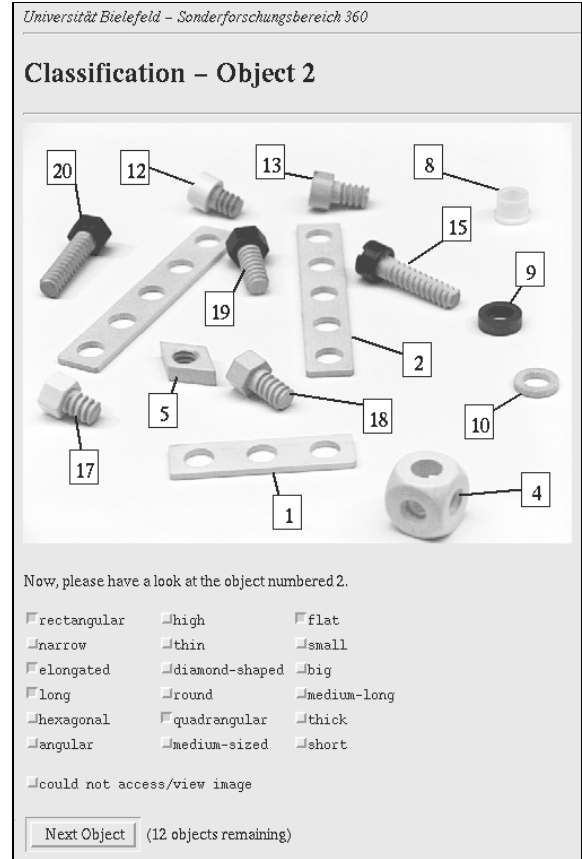


Fig. 7: One page from the questionnaire in the World Wide Web: The objects are presented in the context of others. The subjects were asked to select all size and shape categories that characterize the numbered object.

The main hypotheses or questions for the design of the questionnaire were:

- (a) there is an intrinsic size and shape model for each object in our domain,
- (b) the size and shape depend on the context of the scene, but how?

After an introduction, a WWW page for each object of the domain was presented to subjects. On each page, an image of an object and buttons with size and shape categories were shown, and the subjects were asked to select all those categories that characterize the depicted object (see Fig. 7 as an example). We designed two different versions of the questionnaire according to our hypotheses. In the first version, we presented images of objects in the context of others. The object in question was marked with a number. The second version contained only images of isolated objects. Both versions were randomly distributed among the subjects.

For each object, the subjects could choose from the 18 size and shape adjectives: *small, big, short, long, medium-long, medium-sized, thick, thin, narrow, high, round, angular, elongated, hexagonal, quadrangular, diamond-shaped, flat, rectangular*. We selected the German adjectives from psycholinguistic studies and translated them into English to open this study to international participants. We did not investigate the use of size and shape adjectives in English, and we purely translated the German adjectives according to a dictionary. Thus, the usage of certain adjectives might be different in English and German.

426 subjects with over 20 different native languages completed the questionnaire. The German version was selected by 274 subjects. 96% of them are native speakers. The English version was chosen by 152 subjects, 53% of them are native English speakers. We analyzed the data with descriptive statistics and a χ^2 -test.

Analysis

Fig. 8 shows the means (or relative frequencies) for all size and shape categories for all objects. We use a visualization by greyscale values. The darker the square the greater the value of the mean. Each row corresponds to one category. In a row, there are four squares per object. These stand for the relative frequencies in the data sets from the four versions of the questionnaire (German with context, German without context, English with context, and English without context). We see that the choice of applicable categories is similar for the four different versions of the questionnaire.

We examined the answers of the questionnaires in German in greater detail and observed here even greater similarities. We computed the covariance matrices for the four data sets and used the Frobenius norm to compare the covariance matrices. Fig. 9 shows a 3D histogram plot of the Frobenius norms of the covariance matrices. Despite minor

differences we observe a similar pattern for the uncertainty in the selection of the categories. Furthermore, a χ^2 -test confirms that the influence of the scene context is not significant for the choice of size and shape categories in our scenario.

We consider the observed similarity in the data sets as a hint for intrinsic size and shape models for our domain. A Rolls-Royce car is always a big car even when it stands next to a truck. We think that this is true for the Baufix objects in the assembly scenario, too. Furthermore, the relevant context consists not only of the visible objects but also of personal experience, background, association, etc. The Rolls-Royce is a big car in the context of the knowledge about cars and other car types.

We use the data from the context version of the German questionnaire (137 questionnaires) to estimate the conditional probability tables for our Bayesian network for object identification as described in Subsection 4.2.3. We consider all scene contexts in the same way which leads to satisfying results.

Discussion

Our way of modeling the size and shape of objects models does not attempt to model cognitive processes. So far, we model only one aspect of shape and size perception, which can be captured by an intrinsic size and shape model. But, obviously there must be cognitive processes which are activated in order to compare objects within a specific context. At the beginning of this subsection we raised the question how the size and shape of an object depends on the context of the scene. We are still unable to answer it.

Our approach leads to convincing results which means that we solved the problem of modeling a tool which is able to interact successfully with humans without trying to explain complex cognitive mechanisms.

5 Results

The Bayesian network approach for object identification has been tested in various ways. We show first some examples and demonstrate how objects are identified taking into account an instruction and the objects in a scene.

An analysis with simulated data is used to evaluate our Bayesian network. We generated scenes and instructions randomly and analyze the identification results. This is followed by experiments with real data from psycholinguistic experiments where subjects named specific objects which were presented in images. We compare the identified and the named objects. This section is concluded with an analysis of our approach over time.

5.1 Examples

Our approach is first illustrated with three examples. We take a simple scene with Baufix objects which is depicted

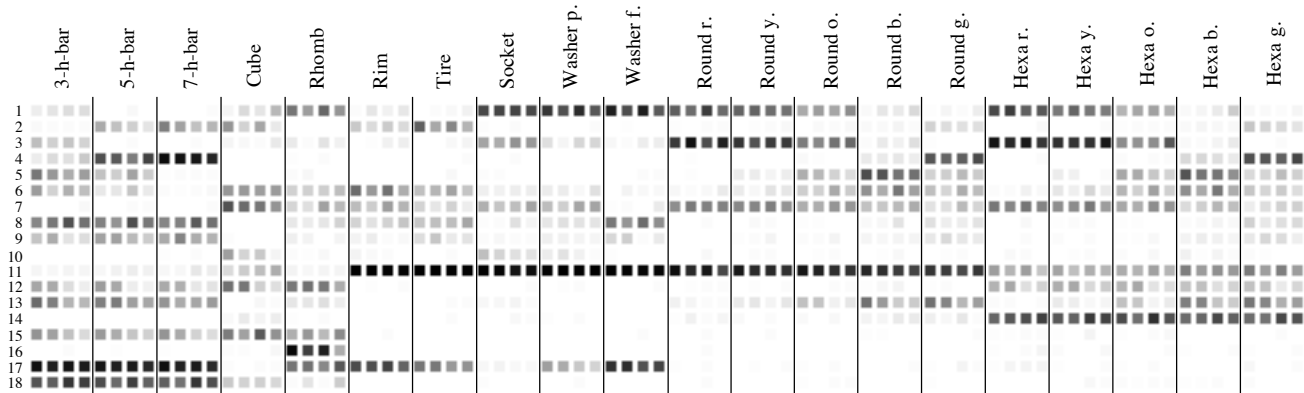


Fig. 8: Relative frequencies for all size and shape categories and all objects in the four versions of the questionnaire: The darker a square the higher the mean or relative frequency of selecting that category. The categories are numbered.

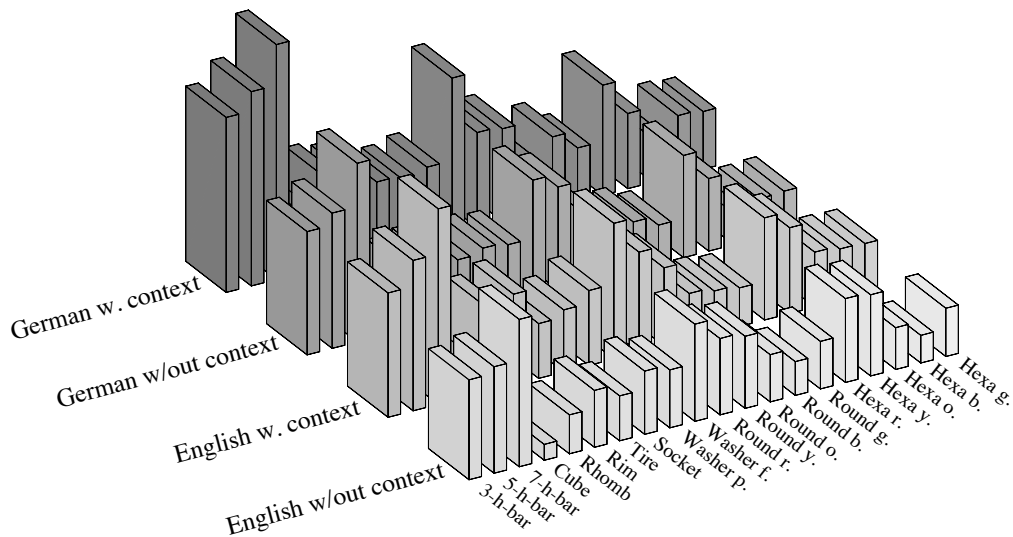


Fig. 9: The Frobenius norm of the covariance matrices for all objects and all versions of the questionnaire.

in Fig. 10. This scene consists of five cubes (two blue, one red, yellow, and green cube), a 3-holed-bar, and a rhomb-nut.

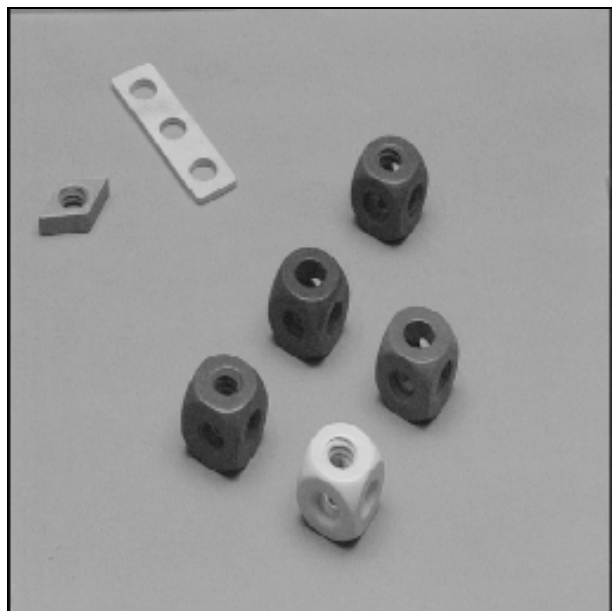


Fig. 10: A scene.

First, we use only information about the scene as input data. This means that we instantiate only the nodes for the objects in the image (\mathbf{object}_i) and propagate the evidence through the network. The resulting beliefs are shown in Fig. 11. Fig. 11 shows the situation before any instruction is given. The object nodes (\mathbf{object}_i) are not depicted here. We see that beliefs in the node **scene** for the categories of objects that occur in the scene are higher than for objects which are not there. The belief for the object category *rim* is rather high because of a high confusion probability between the red rim and the red cube. The beliefs of the node **identified object** are identical to those of the node **scene** as no incoming evidence from the **instruction**³ node is available. The propagated beliefs in the *instruction* nodes report the constellation of the objects in the scene. Most of the objects are cubes and therefore the belief for *cube* is the highest for the property ‘type’. The belief for *bolt* is high as well. This is due to the fact that most of the objects in the Baufix domain are bolts. Hence, the type *bolt* has a high prior probability. This example illustrates that the propagated beliefs are joint probabilities of observed evidence and modeled knowledge.

Fig. 12 is a screen-shot of the object identification result using again the scene shown in Fig. 10 but now including an instruction which names the the categories *object* and *blue*. In the node **identified object**, the beliefs for all blue

objects are higher than others. The belief for the blue cube is the highest because an object of this category is part of the scene. The beliefs for the other blue objects in the node **scene** are due to propagation within the network. The blue cube is clearly the identified object. In this scene, there are two blue cubes (see Fig. 10), which are both identified. The identification result is shown by the system through highlighting the contours of the identified objects in the image of the scene. The user also gets audio feedback with a synthetic voice naming the number and the type of the identified objects.

The speech recognition is done with Hidden Markov Models and a semantic network approach for the linguistic analysis of the uttered phrase (Fink et al., 1994). The result is mapped into a **qualitative description**. This means that the recognition of the instruction results in probability vectors with high beliefs for the recognized categories and small beliefs for non-recognized categories. The content of an instruction is mapped on a predefined set of ‘type’, ‘color’, ‘size’, and ‘shape’ categories in our system.

The third example (Fig. 13) shows the object identification result for the scene shown in Fig. 10 and a more detailed instruction which specifies *cube* and *blue*. Here again the two blue cubes are identified. The belief distribution is similar to that in Fig. 12. However, here the belief for the object category *cube_blue* is even higher due to the unambiguous description. In the node \mathbf{type}_{inst} , the dominating belief is the belief for *cube*, whereas in the node \mathbf{color}_{inst} , small beliefs for the colors *red*, *yellow*, and *green* can still be observed. This is due to the fact that there are cubes of all four colors in the scene. This demonstrates well the interaction of evidence from speech and image data in the network.

In the two latter examples, the two blue cubes are identified. The system exactly responds to the instruction and does not use further selection criteria. In order to constrain the identification to only one object, a more specific instruction is necessary. Spatial relations, for example, “the blue cube *behind* the red one”, can be used to enable the selection of only one specific object.

Our system is implemented in C on a cluster of DEC alpha workstations. A customized framework is used for the inter-process communication. The processing time including the image and speech understanding as well as the object identification is in the range of one minute.

5.2 Simulated Data

A set of two experiments with simulated data was carried out to evaluate the Bayesian network more generally than with single trials. We randomly created 1000 scenes with Baufix objects. For each object and each scene, we tossed a coin whether this object belongs to the scene or not. First, we used the same 23 instructions, each describ-

³This node is not depicted in Fig. 11.

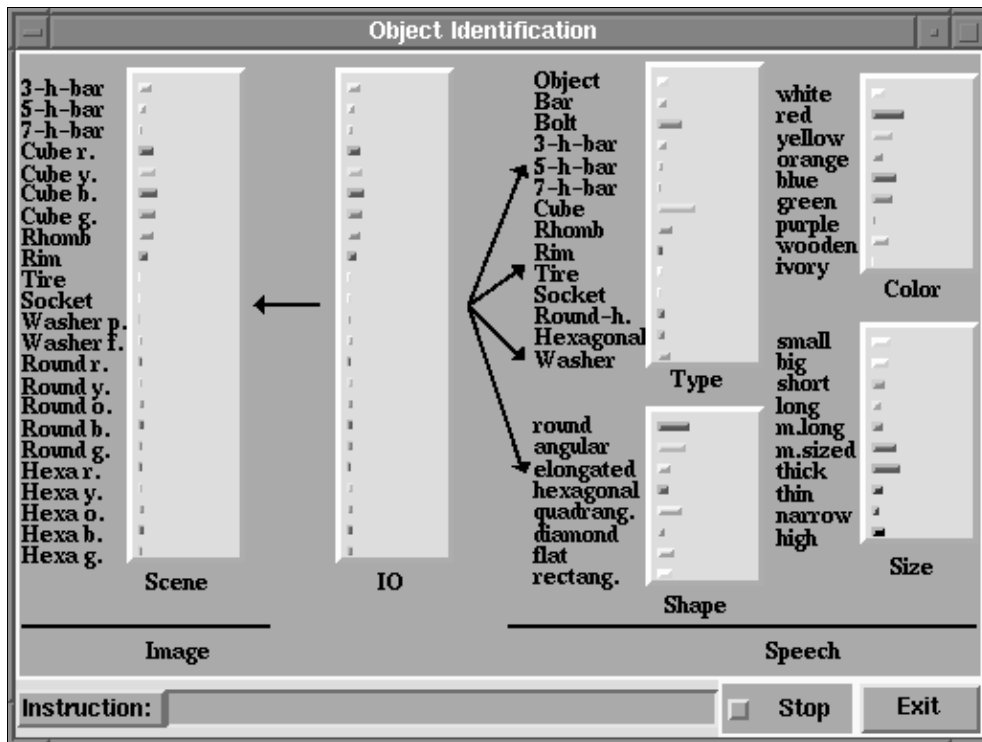


Fig. 11: The belief distribution in the Bayesian network without an instruction, i.e. before any instruction is given.

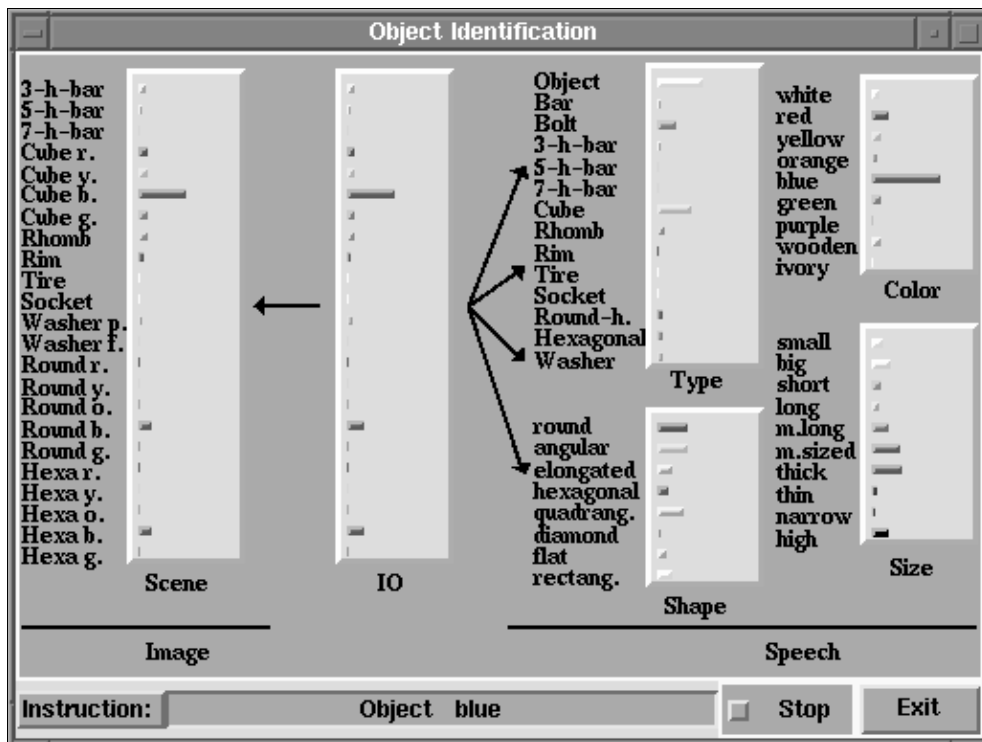


Fig. 12: The belief distribution in the Bayesian network after specifying the categories *object* and *blue* in an instruction: The beliefs are shown after a complete bottom-up and top-down propagation of all evidence.

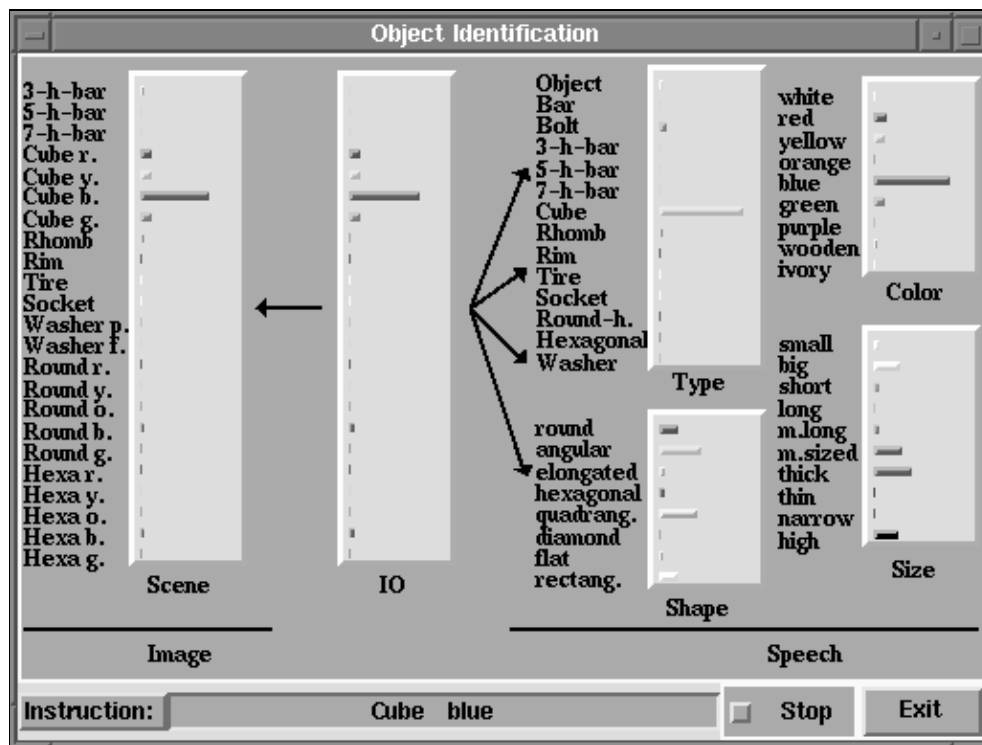


Fig. 13: The belief distribution in the Bayesian network after specifying the categories *cube* and *blue* in an instruction: The belief distribution is shown after the propagation of all evidence.

ing one object class uniquely, for all 1000 scenes. Each instruction describes one object uniquely, but they are more or less specific.

The objects were very well identified. The first two lines of Table 1 report the results. All objects except the socket and the red, round bolt were always uniquely identified when they were present in a scene. The problem for the socket is that its color is not well defined. It is made of some dark white plastic and the reflections caused by this material irritate the color classification. The color is in between white and wooden but obviously not enough clearly separable that it can be well classified in an own class of color ivory. Therefore sometimes the socket is detected as white and sometimes as wooden. Furthermore, the object recognition as a whole is not very stable for this object. The confusions are modeled in the conditional probability tables. However, the object identification prefers well recognized objects and colors.

The other object which was not always uniquely identified is the red, round bolt. This is due to the fact that the categories the used instruction are not specific enough for this object. This shows that the less specific an instruction the smaller is the chance to identify an object.

Table 1 describes the identification results if the object which is intended in an instruction is part of the scene. But,

it is also of interest so see what happens if the intended object is not there. This means that the set of properties which are specified do not apply to any of the objects in the scene. These results can be found in (Socher, 1997). Confusions occur only between objects which have certain properties in common, for example, the red cube and the rim have the same color.

For a larger, second experiment, we randomly generated 200 instructions. These instructions contain for each of the four property classes (type, color, size, and shape) one or no specification. It is obvious that not all of these randomly generated instructions make sense, and sometimes it is hard to distinguish which object is intended. Therefore, two semi-naive⁴ subjects independently classified the meaningful instructions and provided a reference to the object which is in their opinion denoted by the instruction. Those instructions where the subjects agreed on the same object are considered as intending the respective object. 88 instructions were classified as intending a Baufix object. All others are considered as nonsense.

The selected 88 randomly generated instructions were used for all 1000 scenes. We again distinguish the identification results whether the intended object(s) is/are part of

⁴Subjects who are familiar with Baufix objects but not with our object identification approach.

	3-h-bar	5-h-bar	7-h-bar	Cube red	Cube yellow	Cube blue	Cube green	Rhomb-nut	Rim	Tire	Socket	Washer purple	Washer flat	Round red	Round yellow	Round orange	Round blue	Round green	Hexa red	Hexa yellow	Hexa orange	Hexa blue	Hexa green
easy: correct	1	1	1	1	1	1	1	1	1	1	.43	1	1	.69	1	1	1	1	1	1	1	1	1
easy: other	0	0	0	0	0	0	0	0	0	0	.66	0	0	.21	0	0	0	0	0	0	0	0	0
random: correct	.65	.79	.93	.87	.84	.85	.55	.55	.74	.42	.63	0	0	.71	.28	.53	.97	.96	.60	.18	.56	.77	.63
random: other	.38	.16	.06	.34	0	.36	.38	.61	.27	1.21	.86	0	0	.60	.92	.47	.07	0	.82	.75	.59	.29	.20

Table 1: Line 1 and 3: Relative frequency of correct object identifications from the 23 easy descriptions or the 200 random descriptions, respectively. All descriptions were applied to 1000 randomly generated scenes. Line 2 and 4: Relative frequency of additionally or wrongly identified objects when the intended object was present in the scene. These results are obtained when applying the easy and the random descriptions, respectively. A relative frequency greater than 1 indicates that multiple confusions may occur.

the scene or not. The relative frequency of correct identifications per object are shown in line 3 of Table 1. These results apply when the intended object is part of the scene. The relative frequency of additionally or wrongly identified objects is reported in line 4 of Table 1. Here, a relative frequency greater than 1 indicates that multiple confusions occur. Instructions which are intending the two washers were not contained in the randomly generated test set. The results are not as good as for the easy descriptions but still reasonably good. For the three worst cases there are easy explanations. The randomly generated instructions do not characterize the objects well enough.

Again, we checked also the case if the intended object is not part of the scene (Socher, 1997). The range of confusions is wider than for the “easy descriptions”, which is very much due to the fact that the randomly generated descriptions are not consistent within the named properties. Not all specified properties really apply to the intended object.

The two examples, the easy and the randomly generated descriptions, can be considered as a good and a bad example. They show the range of results which are obtained with our Bayesian network approach for object identification.

5.3 Real Data

We also carried out experiments with real data. In the data set described in Subsection 4.2.3, subjects referred to objects which were shown on a computer screen. We used this data set in three ways. (1) as it is, (2) we transcribed the instructions orthographically and used textual input for our experiment instead of speech, and (3) we ‘idealized’ the real data. This means that we use the orthographic transcription of the instructions and manually generated scene descriptions instead of images in order to evaluate only the Bayesian network approach. Errors in the understanding processes are avoided.

The Bayesian network approach is evaluated by comparing the object identification results with the originally

referred, intended objects⁵. Unfortunately, we also used these data for the estimation of the conditional probability tables in the Bayesian network, and due to time constraints we were not able to acquire another test set.

An identification is *correct* if the intended object is found. If additional objects were identified besides the intended object then it is still considered as correct but counted for the class *additional* as well.

The best identification results are achieved with the ‘idealized data’. Errors in speech recognition have the greatest impact on the overall results. If an instruction is at least partially understood, then more than 70% of the objects are correctly identified. When we avoid recognition errors, then the intended object is correctly identified for 92.5% of the instructions. False or additional identifications occur here mainly because of inaccurate or imprecise specifications by the subjects. Another reason for errors (*false/additional*) is that the identification criterion ($\eta_j > \mu + \sigma$) is not adequate for a given instruction and scene. It is a dynamic threshold, but thresholds may fail.

5.4 Spatial Relations

We collected a second set of instructions under the same conditions as described in the previous paragraph or in Subsection 4.2.3. This set is used only for testing. Six subjects had to name marked objects from five different scenes which were presented on a computer screen. This time, the subjects were explicitly asked to use spatial relations in every instruction. We ‘idealized’ the data again to evaluate the performance of our identification approach when spatial relations are used. The evaluation is carried out in the same way as described in the previous paragraph. In addition to that, we require for this experiment that the object which is identified by the system must be exactly the same, i.e. having the same image coordinates as the one which was named by a subject. It is possible to specify an object

⁵The subjects described in each instruction one object which was marked in an image by an arrow.

source	# instructions	<i>correct</i>	<i>additional</i>	<i>false</i>	<i>nothing</i>
idealized	412	381 (92.5%)	34 (8.3%)	31 (7.5%)	0 (0%)
text	417	360 (86.3%)	34 (8.1%)	40 (9.6%)	17 (4%)
speech	133	93 (70%)	14 (10.5%)	25 (18.8%)	15 (11.2%)

Table 2: Object identification results using ‘idealized data’, textual, and speech input: *Correct* identifications (may contain additional objects besides the intended object), identifications with *additional* objects, *false* identifications, and cases where no object was identified (*nothing*).

uniquely with spatial relations which wasn’t the case with only the description of the type, color, size, or shape of an object. The results are reported in Table 3.

The number of false identifications is now higher than for the experiment shown in Table 2. This is due to the more severe criterion for a correct identification, but also because of discrepancies between the computation of spatial relations and the use of spatial relations by the subjects (Vorwerk et al., 1997).

5.5 Image Sequences

In our targeted applications, the scenes are not changing very much. The assembly is rather slow compared to the image frame rate. Furthermore, no actions in the scene might happen during the time an instruction is uttered. Therefore, changes in the object hypotheses are more likely due to recognition errors than to changes in the scene. This is especially true when the image region characteristics (i.e. center of mass), which correspond to the recognized object, are basically constant in subsequent image frames.

We can compensate these errors easily with our Bayesian network. We simply use the beliefs computed for the object nodes ($\mathbf{object}_1, \dots, \mathbf{object}_m$) at time $t - 1$ as causal support for these nodes at time t . Thus, we predict that the object category will not change. Fig. 14 shows the results of two experiments with simulated data. Both experiments run over 80 time frames. One object node is observed during this time. Random noise is added to the object hypotheses. The probability that the object hypothesis does change is set to 0.25. Hence, the object hypothesis may vary in a quarter of all time frames. Fig. 14 shows four components of the belief vector of the object node. The true object category is represented by one of these four components. Fig. 14a shows the results when recursive prediction is active. We see the noise in the data, but the belief of the true object hypothesis is clearly distinguishable from the others. In Fig. 14b the results are shown of the same experiment but without prediction. The signal to noise ratio is here almost equal to 1. The noise distorts the the belief of the true object hypothesis significantly.

6 Discussion and Conclusions

We described a representation for image and speech understanding results and its application in an integrated im-

age and speech understanding system for natural human-computer interaction. This representation can also be used for other types of qualitative features and other domains. The vectorial representation has the advantage of being suitable for multiple and even contradicting results, as well as for overlapping categories. A variety of information can be represented which does not require hard decisions to be made at early stages of understanding processes. This form of representation has been proven to work successfully in our system, and it forms the basis for the integration of image and speech understanding. It lends itself for probabilistic reasoning upon the represented entities and is not specifically adapted to specific underlying computational routines.

Bayesian networks provide a formalism for reasoning about partial beliefs under conditions of uncertainty. They are very well suited for integrating different extracted properties and weigh the influence of them upon the final object identification accounting for uncertainties in the data, the detection processes, and the decision process. Empirical data and results from psycholinguistic experiments can be easily incorporated. Our object identification approach is rather flexible and well suitable for the variety of different instructions humans can give to the system.

Comparing related work, we find that representations used in most computational systems are rather simple and specifically adapted to the actual task (e.g., André et al., 1988; Kollnig & Nagel, 1993; Tsotsos et al., 1997). The goal is, in most cases, to extract specific symbols from images. Reasoning on these representations or inferring other results from them is often not considered. Whereas typical reasoning approaches do not start from sensory data in most cases but already from high-level special purpose representations (e.g. Egenhofer, 1991; Hernández, 1993).

In the discussion of our approach, we observe the following items:

- The Bayesian network approach is a decision calculus. It determines the object with the highest joint probability of the named and detected features. Knowledge is modeled mainly through the conditional probability tables. No generalizing or reasoning rules are incorporated. The system is not able to

source	# instructions	correct	additional	false	nothing
idealized	98	82 (83.5%)	2 (2%)	16 (16.5%)	0 (0%)
text	84	66 (78.6%)	7 (8.3%)	18 (21.4%)	0 (0%)
speech	21	16 (76%)	12 (57%)	5 (24%)	0 (0%)

Table 3: Object identification results for instructions with spatial relations: *Correct* identifications (may contain additional objects besides the intended object), identifications with *additional* objects, *false* identifications, and cases where no object was identified (*nothing*).

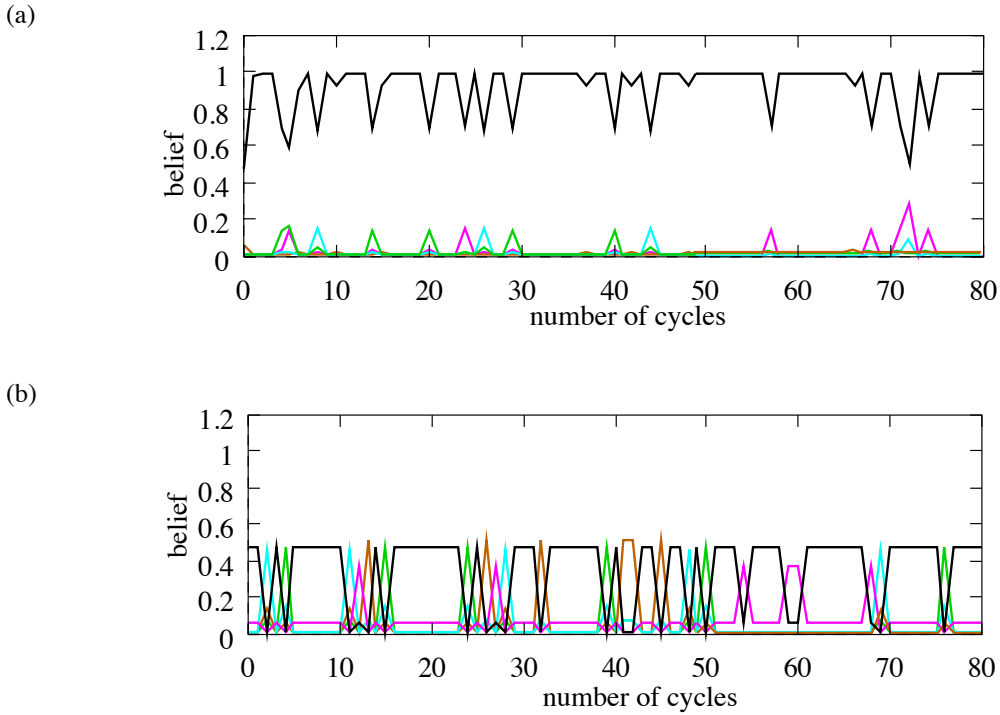


Fig. 14: The behavior of the beliefs in an object node over time: Four components of the belief vector are plotted in different greyvalues. White noise was added to the diagnostic support. (a) With prediction: The belief of the true object hypothesis is always greater than the others, (b) Without prediction: the signal to noise ratio is close to 1.

generalize or to abstract from named properties. No hierarchies like

$$\begin{aligned}
 &(\{\text{rectangular, diamond-shaped}\} \in \text{quadrangular}) \\
 &\cup \{\text{hexagonal}\} \in \{\text{angular}\}
 \end{aligned}$$

are implemented.

- The system is more likely to identify an object than not. If there is at least one object property which matches an uttered property then the object will be identified if no other object fits better. It can be assumed that a serious instructor does not want to fool the system. Instructions should refer to objects in the scene, and they somehow should make sense. In this case, it is a good strategy for the system to identify the best matching object without requiring a perfect match. Furthermore, it is more likely for the system
- to identify only the best fitting object rather than all possible fits.
- The Bayesian network approach takes the scene context into account. However, no topological structure is considered. Local neighborhoods of objects or sub-scenes can not be represented. Therefore, an explicit focus on an object and its local neighborhood can not be modeled nor recognized. A possibility could be to model the scene topology with Markov Random Fields (Geman & Geman, 1984) and to attach them to the nodes in the Bayesian networks. Then, evidence would not be derived from the qualitative descriptions only, but from the qualitative descriptions within the context of the scene topology.
- The identification criterion ($\eta_j > \mu + \sigma$) is a dynamic threshold. As for all hard decisions, also this thresh-

old may fail. Future work should investigate learned identification criteria, for example, neural networks or other classifiers.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) in the project SFB 360 and the German Academic Exchange Service (DAAD) under the grant program HSP II/AUFE. Collaborations with Constanze Vorwerg, Thomas Fuhr, and Franz Kummert have been very fruitful for this work.

References

- André, E., Herzog, G., & Rist, T. (1988). On the simultaneous interpretation of real world image sequences and their natural language description: The system SOCCER. In *Proc. of the 8th European Conference on Artificial Intelligence (ECAI-88)*, pp. 449–545.
- Egenhofer, M. J. (1991). Reasoning about binary topological relations. In O. Günther & H.-J. Schek (Eds.), *Advances in Spatial Databases, 2nd Symposium, SSD'91*, Berlin, pp. 143–157. Springer.
- Fink, G. A., Kummert, F., & Sagerer, G. (1994). A Close High-Level Interaction Scheme for Recognition and Interpretation of Speech. In *Proc. Int. Conf. on Spoken Language Processing*, Volume 4, Yokohama, Japan, pp. 2183–2186.
- Fuhr, T., Kummert, F., Posch, S., & Sagerer, G. (1993). An Approach for Qualitatively Predicting Relations from Relations. In E. Sandewall & C. G. Jansson (Eds.), *Proc. of the Scandinavian Conference on Artificial Intelligence*, Amsterdam, pp. 38–49. IOS Press.
- Fuhr, T., Socher, G., Scheering, C., & Sagerer, G. (1997). A three-dimensional spatial model for the interpretation of image data. In P. Olivier & K.-P. Gapp (Eds.), *Representation and Processing of Spatial Expressions*. Hillsdale: Lawrence Erlbaum Associates. To appear.
- Geman, S. & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. on Pattern Recognition and Machine Intelligence PAMI-6*, 721–741.
- Harnad, S. (1987). Introduction: Psychophysical and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception. The groundwork for cognition*, pp. 1–25. Cambridge University Press.
- Heidemann, G., Kummert, F., Ritter, H., & Sagerer, G. (1996). A Hybrid Object Recognition Architecture. In *International Conference on Artificial Neural Networks ICANN-96*, Bochum, Germany, July 15-19.
- Hernández, D. (1993). *Qualitative Representation of Spatial Knowledge*. Lecture Notes in Artificial Intelligence, 804. Berlin, Heidelberg, etc.: Springer-Verlag.
- Herrmann, T. & Deutsch, W. (1976). *Psychologie der Objektbenennung*. Bern: Huber.
- Koller, D., Daniilidis, K., & Nagel, H.-H. (1993). Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes. *International Journal of Computer Vision 10*(3), 257–281.
- Kollnig, H. & Nagel, H.-H. (1993). Ermittlung von begrifflichen Beschreibungen von Geschehen in Straßenverkehrsszenen mit Hilfe unscharfer Mengen. *Informatik Forschung und Entwicklung 8*, 186–196. (in German).
- Mc Kevitt, P. (Ed.) (1994). *Special Issue on Integration of Natural Language and Vision Processing*, Volume 8 of *Artificial Intelligence Volume*. Kluwer Academic Publishers.
- Medin, D. L. & Barsalou, L. W. (1987). Categorization processes and categorical perception. In S. Harnad (Ed.), *Categorical perception. The groundwork for cognition*, pp. 455–490. Cambridge University Press.
- Nagel, H. (1988). From image sequences towards conceptual descriptions. *Image and Vision Computing 6*(2), 59–74.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers.
- Russell, S. & Norvig, P. (1995). *Artificial Intelligence A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.
- Socher, G. (1997). *Qualitative Scene Descriptions from Images for Integrated Speech and Image Understanding*. Dissertationen zur Künstlichen Intelligenz (DISKI 170). Sankt Augustin: infix-Verlag.
- Srihari, R. K. (1994). Photo Understanding Using Visual Constraints Generated from Accompanying Text. In P. Mc Kevitt (Ed.), *AAAI-94 Workshop on Integration of Natural Language and Vision Processing*, pp. 22–29. Twelfth National Conference on Artificial Intelligence (AAAI-94).
- Toal, A. F. & Buxton, H. (1992). Spatio-temporal reasoning within a traffic surveillance system. In *Proc. Second European Conference on Computer Vision*, pp. 884–892. Santa Margherita Ligure, Italy, 18-23 May, G. Sandini (Ed.), Lecture Notes in Computer Science 588, Springer-Verlag.
- Tsotsos, J. K., Verghese, G., Dickinson, S., Jenkin, M., Jepson, A., Milios, E., Nuflo, F., Stevenson, S., Black, M., Metaxas, D., Culhane, S., Yet, Y., & Mann, R. (1997). PLATBOT: A Visually-Guided Robot for Physically Disabled Children. submitted to *Image and Vision Computing*.
- Vorwerg, C., Socher, G., Fuhr, T., Sagerer, G., & Rickheit, G. (1997). Projective relations for 3D space: Computational model, application, and psychological evaluation. In *Proc. of the 14th National Conference on Artificial Intelligence (AAAI-97)*, Providence, Rhode Island, pp. 159–164.
- Wahlster, W. (1989). One Word says More Than a Thousand Pictures. On the Automatic Verbalization of the Results of Image Sequence Analysis Systems. *Computers and Artificial Intelligence 8*, 479–492.