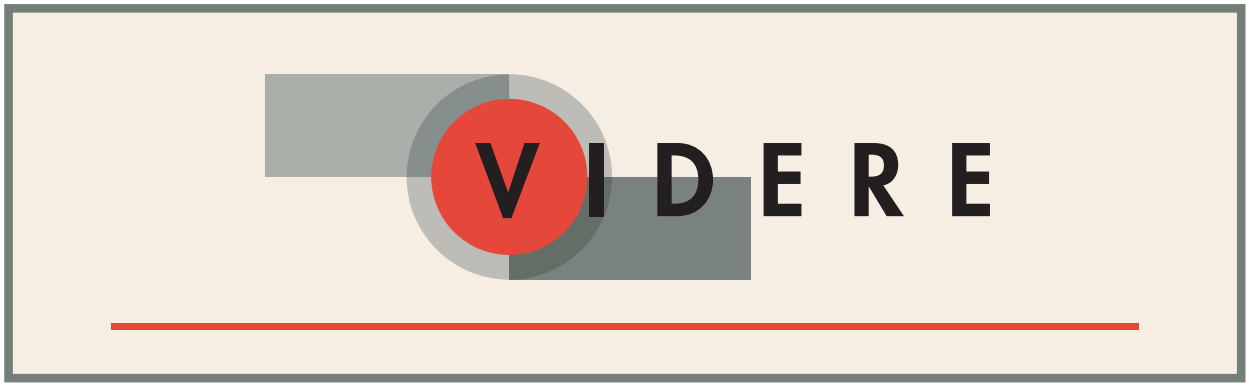


Videre: Journal of Computer Vision Research

Quarterly Journal
Winter 2000, Volume 1, Number 4
The MIT Press

Article 3

Integration of Vision and Speech Understanding Using Bayesian Networks

Sven Wachsmuth
Gudrun Socher
Hans Brandt-Pook
Franz Kummert
Gerhard Sagerer

Videre: Journal of Computer Vision Research (ISSN 1089-2788) is a quarterly journal published electronically on the Internet by The MIT Press, Cambridge, Massachusetts, 02142. Subscriptions and address changes should be addressed to MIT Press Journals, Five Cambridge Center, Cambridge, MA 02142; phone: (617) 253-2889; fax: (617) 577-1545; e-mail: journals-orders@mit.edu. Subscription rates are: Individuals $30.00, Institutions $125.00. Canadians add additional 7% GST. Prices subject to change without notice.

Integration of Vision and Speech Understanding Using Bayesian Networks

Sven Wachsmuth[1], Gudrun Socher[2], Hans Brandt-Pook[3], Franz Kummert[3], Gerhard Sagerer[3]

The interaction of image and speech processing is a crucial property of multimedia systems. Classical systems using inferences on pure qualitative high-level descriptions miss much information when concerned with erroneous, vague, or incomplete data. We propose a new architecture that integrates various levels of processing by using multiple representations of the visually observed scene. The representations are vertically connected by Bayesian networks in order to find the most plausible interpretation of the scene.

The interpretation of a spoken utterance naming an object in the visually observed scene is modeled as another partial representation of the scene. Using this concept, the key problem is the identification of the verbally specified object instances in the visually observed scene. Therefore, a Bayesian network is generated dynamically from the spoken utterance and the visual scene representation.

Keywords: integration of speech and vision, vision, speech understanding, correspondence problem, multi-modal input, spatial modeling, Bayes nets, graph matching.

1. swachsmu@techfak.uni-bielefeld.de

2. Online Anywhere, 3145 Porter Dr., Bldg. A #202, Palo Alto, CA 94304

3. University of Bielefeld, Technical Faculty, P.O. Box 100131, 33501 Bielefeld, Germany

1 Introduction

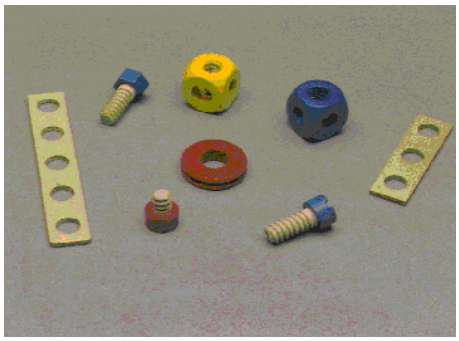
Human-machine interfaces are one of the major bottlenecks when using computer systems in real environments. One possibility to overcome these drawbacks is using multiple modalities in the communication between humans and machines, which is a quite natural concept for human communications. In this case, we distinguish different perceptive channels by which the system is connected with the environment. The consequence is that we have to generate a common interpretation of the transmitted information on all channels or modalities instead of analyzing them in an isolated way. In order to combine different perceptive channels, they have to be connected by an internal knowledge representation. In this paper, we concentrate on the integration of image and speech processing in a situated environment.

We developed a system for the following concrete scenario. A human has to instruct a robot in a construction task. Therefore, both communication partners perceive an arrangement of building parts on a table (figure 1). These can be assembled by screwing or plugging. The human speaker has to verbally specify the objects in the scene by describing properties of these objects without knowing the exact terms. Therefore, the system has to interpret mostly vague descriptions. The system is capable of viewing the objects in the scene and has a speech-recognition and speech-understanding unit. Because we use our system in a real environment and it has to interpret vague descriptions, the mapping of the verbal object descriptions and the visually perceived objects is a very tough task. In figure 1, we present an example dialogue that may happen in our domain. The user views the scene on a table and instructs the robot which object it should grasp.

A rather easy way to combine results of the vision- and speech-understanding processes is shown in figure 2a. Both processes analyze their input signals separately through different levels of abstraction and generate some high-level description using common logical predicates. Afterwards, logical inferences can be drawn to get a consistent interpretation of the input data. When applying such an approach in noisy environments, any decision in lower processing levels results in a loss of information on higher levels. Thereby, logic inferences on the top level of description may fail even when the separate interpretations are only slightly incorrect.

We propose a new architecture (figure 2b) that integrates multimodal processing on several levels of abstraction. A careful approach to uncertainty is needed for a consistent interpretation of the input data that takes into account the several layers and the dependencies between

Figure 1. Example dialogue of the domain.



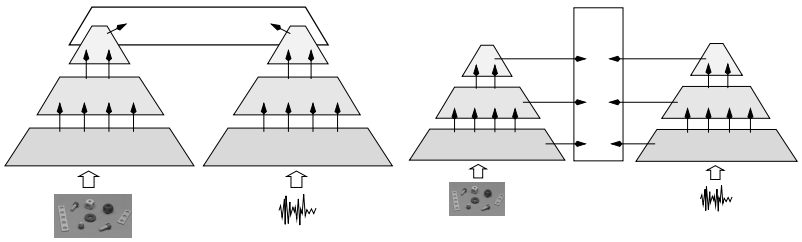
user: "Take the bar."

system: "I have found two bars. Should I take the long one or the short one?"

user: "Take the long one."

system: "O.k., I will take the five-holed-bar."

Figure 2. Integration of vision and speech understanding.



(a) High-level integration (b) Multi-level integration

them. Therefore, we use Bayesian networks as a decision calculus that can handle uncertainties, erroneous data, and vague meanings in a very comfortable way.

Many people who were asked to identify an object in our domain used spatial relations in order to specify the location of the intended object. Therefore, an important issue in our system is the modeling of space. Based on the definition of projective relations between two objects in three dimensions, we show how to approximate these when data is available only in two dimensions and how to integrate context information about the topology of the presented scene.

In section 2, we will briefly review some related approaches dealing with the problem of vision and speech integration. Then we will describe some important aspects of the image, and speech-processing components (section 3) and the spatial model (section 4) in our system. Afterwards, we will propose a new approach that integrates these two modalities using multiple representation layers (section 5). We emphasize that the spatial modeling is an important aspect in this interaction scheme. Finally, we will give some experimental results of the implemented system (section 6) showing the robustness of our approach and a short conclusion (section 7).

2 Related Work

In literature the topic of integrating vision and speech understanding is referenced from different viewpoints [22]. The construction of mental pictures [8] can be induced by verbal descriptions or previously seen objects. They are used to reason about scenes which are currently not visible. This is an important aspect in language understanding when spatial knowledge is involved [1, 17, 30]. Other systems [7, 14, 29] try to generate verbal descriptions from images or image sequences. They realize an automatic generation of qualitative representations from image data that is fundamental for integration of vision and speech understanding and use various approaches to modeling spatial relations. Lastly, much work has incorporated both linguistic and pictorial inputs concerning

the interpretation of textual annotated pictures [23], lip reading [2, 10], or multimedia systems [13]. In the following, we concentrate on systems that visually observe the scene to enable natural human-computer interaction.

The PLAYBOT project [24, 25] was started to provide a controllable robot that might enable physically disabled children to access and manipulate toys. The robot possesses a robotic arm with a hand, a stereo color vision robot head, and a communication panel. The child gives commands on the communication panel which displays actions, objects, and locations of the objects. The command language consists of verbs, nouns, and spatial prepositions. The child selects the "words" by pointing on the communication panel. While language processing is simplified by using the communication panel, most attention is given to processing visual data. Object recognition [4] is performed by fitting deformable models to image contours. Object tracking [26] is attained by perspective alignment. The event perception [18] is based on an ontology suitable for describing object properties and the generation and transfer of forces in the scene. To find instances of a target object in the image, a Bayesian network approach is employed which exploits the probabilities in the aspect hierarchy of modeled objects which is used in object recognition.

The ubiquitous talker [16] was developed to provide its user with some information related to a recognized object in the environment. The system consists of an LCD display that reflects the scene at which the user is looking as if it were a transparent glass, a CCD camera for recognizing real-world objects with color-bar ID codes, and a microphone for recognizing a human voice. The main aspect of the system is the integration of the linguistic and nonlinguistic contexts to interpreting natural language utterances (which becomes much easier when the situation is fixed by nonverbal information). The system basically consists of two subsystems: the subsystem that recognizes a number of real-world situations that include objects with color-bar ID codes, and another subsystem that recognizes and interprets user speech inputs. The image (color-code) recognizer triggers the speech recognizer and sends a message to it to select the appropriate vocabulary and grammar for analyzing the spoken utterance. There are two connections between linguistic and nonlinguistic contexts. User's intentions are abductively inferred by using a plan library [15]. This process is initially triggered by introducing a new nonlinguistic context. Another connection is deictic centers [31] that are possible referents of deictic expressions. The object and the location in a nonlinguistic context can be current deictic centers. The preferences on possible deictic centers as a referent are based on dialogue information.

Two aspects of the mentioned systems are important for the integration of vision and speech understanding in our domain. Firstly, in PLAYBOT, a Bayesian network is used to calculate the most plausible interpretation of an aspect hierarchy that describes possible target objects on different levels. Secondly, the ubiquitous talker uses for speech recognition specific vocabularies and grammars that are selected by the situative context and uses detected objects and locations as referents for deictic expressions. In our domain, we also have a nonlinguistic context because the speaker is confronted with a scene of building parts and has to specify an intended object.

Both mentioned systems simplify the task of matching the interpretation of the visually observed scene and the verbal instruction of the user. Recognized objects are visualized on a screen and are directly accessible via a communication panel (PLAYBOT) or some textual information about them is shown on a display and can directly be referred to by speech (ubiquitous talker). Another difference is that our system has to deal with uncertainties on both modalities, while PLAYBOT is only concerned with uncertainty in vision processing and the ubiquitous talker only with uncertainty in speech recognition.

3 System Components

In order to motivate the interaction scheme of image and speech processing, we will briefly introduce some important aspects of the components of our systems.

3.1 Hybrid Object Recognition

The vision component is based on a hybrid approach [9] that integrates structural and holistic knowledge about the real-world objects in a semantic network formalism. An HSV camera image is segmented into homogeneously colored regions. Every region is characterized by some shape features, and its center point is classified by a special artificial neural network [6]—the local linear map (LLM)—using a 16-dimensional feature vector. By this classification, we get a holistic, 2-D object hypothesis that is verified by the structural knowledge of the semantic net using the features of the colored regions.

Because precise spatial reasoning is possible only in three dimensions and the robot needs the 3-D position and orientation to grasp an object, we calculate a reconstruction in space for all 2-D hypothesis based on simple geometric object models [19, 20]. Geometric features like points, lines, and ellipses are fitted to detected image features [12] using an iterative optimization technique.

3.2 Integrated Speech Recognition and Understanding

The speech component of our system consists of two subcomponents—the recognizer and the understanding component—which are tightly coupled. The understanding component is realized in a semantic network that models both linguistic knowledge and knowledge about the construction task. Both knowledge types are incorporated into the recognition process using an LR(1)-grammar [28]. An important aspect in the interaction of vision and speech is extracting object descriptions from spoken utterances. These are especially modeled by the used grammar so that the understanding component receives structured hypotheses from the recognition component [3]. On a higher level in the understanding component, feature structures are generated specifying an intended object by type, color, size, and shape attributes and by spatial relations using reference objects. Attributes, such as “color:red”, can be instantiated even on word level using a lexicon while instantiating reference objects, such as “ref:{rel:left,{type:cube,color:blue}}”, needs more structural analysis.

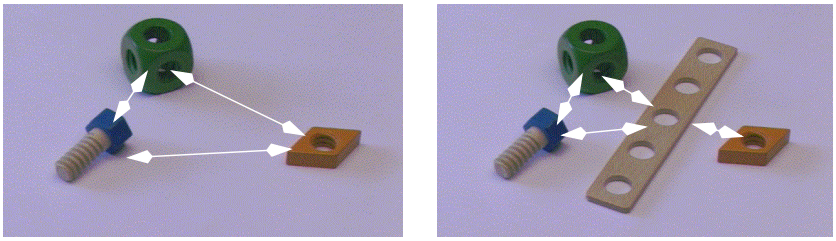
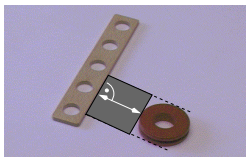
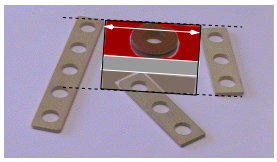


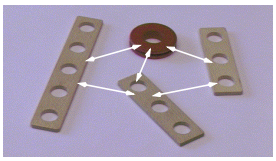
Figure 3. Topological influence of additional objects. If we add a bar between the cube, the bolt, and the rhomb-nut, two neighborhood relations vanish and three other neighborhood relations are newly introduced.



(a) Area between objects.



(b) Degree of 'separation' of the left and right bar.



(c) Neighborhood relations in the scene.

Figure 4. Definition of neighborhood based on object regions. (a) The rectangled area between two objects is defined by the shortest path between the two region boundaries and their expansion in the orthogonal direction. (b) The dark colored sections of the area between the left and right bar define the percentage by which they are separated. (c) The left and right bar are not neighbors because their object regions are separated by regions of two other objects.

4 Spatial Model

Human object localization via spatial relations, such as "take the object left to the long one," is qualitative, vague, and context sensitive. When we try to construct a spatial model of our scene, various aspects are involved. Firstly, there are inherent properties of the two objects mentioned with a binary spatial relation, such as direction, relative distance, orientation, and extension of the reference object [27]. Secondly, we have a large influence of the local configuration of the other objects when considering scenes with collections of objects. The user will name relations between only those objects that are obvious for him. Therefore, an important concept is "neighborhood" and we use it to define the topology of the visible scene. The neighborhood of two objects cannot be defined only by inherent properties of an object pair. Additional objects in the scene can strongly influence this property (figure 3). The concept of neighborhood is basically defined on local separation of objects. Using 3-D data, we model relations between closed object volumes; using 2-D data, we model relations between object regions. (For simplicity, we present only the 2-D model in this section.) But the whole concept can be expanded to 3-D data in a similar way. In order to calculate the 2-D neighborhoods we take the segmented regions from 2-D object recognition. This can be interpreted as an approximation of the local configuration of the objects in the scene, because it is defined by only one camera view.

One object region is "separated" from the other if the area between them is overlapped by other object regions in the scene more than a specified percentage (figure 4). The term *neighbor* is applied to every pair of object regions that are not "separated". When all neighborhoods are combined, we get a neighborhood graph of the scene. This graph will be used when combining the vision results and the speech input.

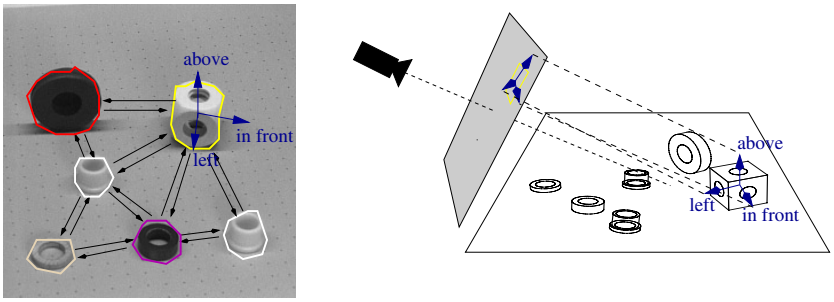


Figure 5. The direction vectors from the 2-D spatial model can be interpreted using the projection of the reference frame of the user to the 2-D image.

4.1 Projective Relations

Because we are living in a 3-D world, precise spatial reasoning can only be accomplished using a 3-D model of space. The system cannot disambiguate all directions if it knows only one 2-D view of the scene. Especially if you consider arbitrary locations of objects (which may not be on the planar table) and overlapping regions of objects, the system needs a 3-D spatial model to resolve the description of an intended object. Nevertheless, we can observe that humans sometimes use projective relations like *above* or *below* in a 2-D manner as they were looking at a picture and do not imagine the 3-D scene [27].

In our domain, projective relations are used to describe the spatial location of an intended object (IO) in the scene relative to another object, the reference object (RO). Which relation is named by the user depends on the currently used reference frame, which can change in every utterance, because the user may look on the scene from different viewpoints. Our computational model for spatial relations is designed to compute the binary relations left, right, above, below, behind, and in front. They are defined on different levels.

On the 2-D level, spatial relations between two objects are represented by a vector that describes the orientation of the shortest path between both regions associated with the objects. In a second step, the reference frame (ref) is projected into the image (figure 5) and a degree of applicability is computed in regard to the angle ϕ between the 2-D vector of the spatial relation and the vector of the mentioned projective relation (rel):

$$App(ref, rel, IO, RO) = \alpha(ref, rel, IO, RO) = 1 - \phi\tilde{u}.$$

The 3-D level is much more complex. The computational model is able to cope with different frames of reference, represents vague and overlapping meanings of projective relations, and considers the influence of the objects' shape on the applicability of prepositions. A detailed description can be found in [5]. The rough idea of the model is presented in figure 6.

Instead of detailed geometric models of the objects, we use surrounding boxes as abstractions that are collinear to the objects' inertia axes (figure 6b). A finite number of acceptance volumes is associated with each object (fig. 6c). These are infinite open polyeders bound to the sides, edges, and corners of the object. They partition the 3-D space surrounding the object. A direction vector corresponds to each acceptance volume. It roughly models the direction to which an acceptance volume extends in space.

The computation of spatial relations from objects is a two-layered process (figure 6d). In the first layer, a reference-independent spatial representation is computed. It can be expressed by a set of acceptance

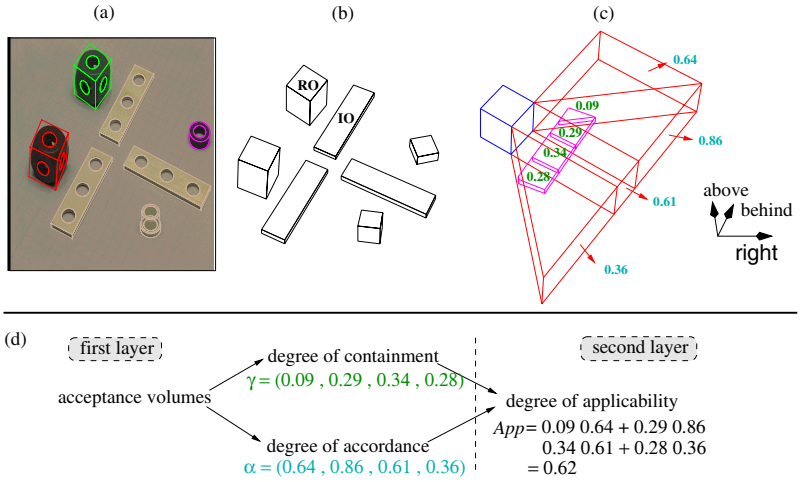


Figure 6. Computation of 3D spatial relations. (a) shows the adapted CAD-models of the recognized objects. In (b) they are approximated by bounding boxes. (c) shows the representation of the first layer concerning the spatial relation (*'right'*) of an object pair. (d) A degree of applicability App can be computed on the second representation layer.

relations that are associated with each acceptance volume. They are judged by the degree of containment $\gamma(IO, RO)$ with regard to the intended object. Also calculated in the first layer are the reference-dependent meaning definitions of relations rel (for example "right") with regard to certain reference objects and a given reference frame ref of the user. The membership of an acceptance relation is judged by its degree of accordance, which is computed using the angle between the direction vector of the acceptance volume and the direction of the relation rel. These two judged 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. We get a degree of applicability App by calculating the scalar product:

$$App(ref, rel, IO, RO) = < \alpha(ref, rel, RO) | \gamma(IO, RO) > .$$

In order to apply this computational model to a spatial relation, both objects must be reconstructed in 3-D space. In many cases, we do not need precise 3-D relations to identify an indented object that was denoted by a spoken utterance. In these cases, we are able to reason about spatial descriptions using the 2-D spatial model even if there were no 3-D information available according to time constraints or recognition or segmentation failures.

4.2 Multilevel Representation of the Scene

In the preceding sections, we described how we model different aspects of the visual scene, like topology and spatial relations. If we want to use this information plus the results from object recognition in an integrated way, we need a comprising representation of these aspects.

For this purpose, we define a labeled graph

$$\mathcal{G}_{vision} = (\mathcal{V}, \mathcal{E}), \quad \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V},$$

which we call *neighborhood graph* because the edges in this graph represent neighborhood relations between objects. The nodes $v \in \mathcal{V}$ are labeled with object hypotheses $Obj(v)$ from vision. The edges $e = \{v_1, v_2\} \in \mathcal{E}$ are labeled with reference-independent representations of the 2-D and 3-D spatial relations $Rel(Obj(v_1), Obj(v_2))$, $Rel(Obj(v_2), Obj(v_1))$.

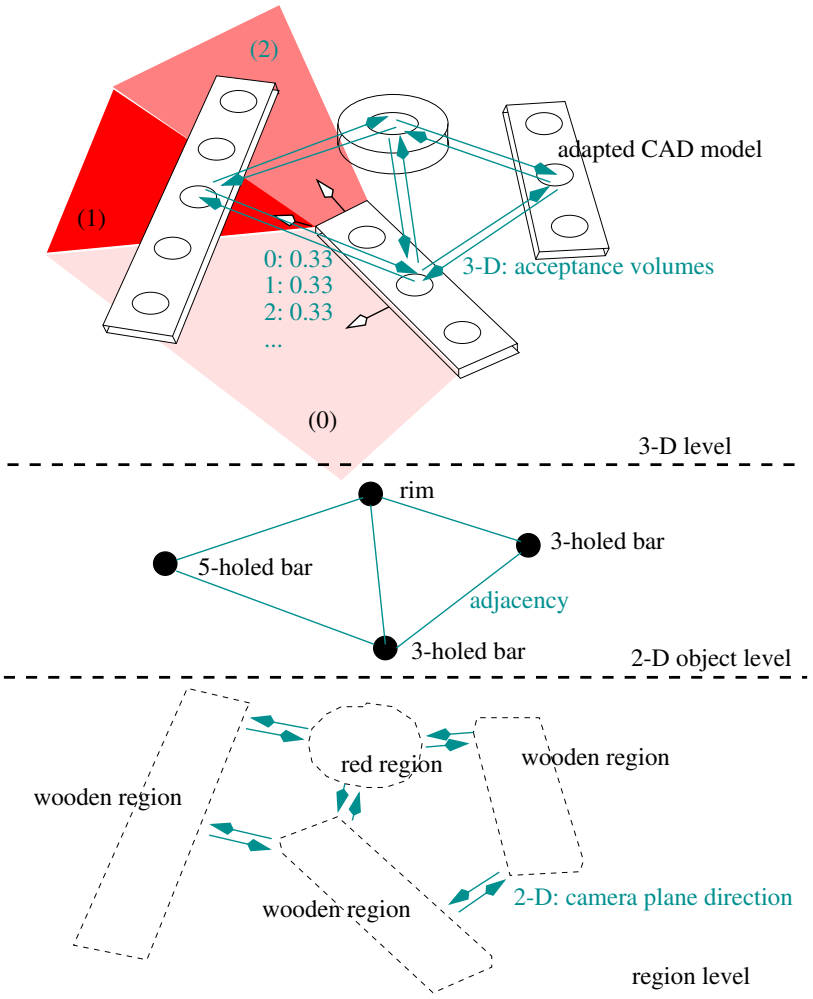


Figure 7. Multilevel representation of the scene.

The labels of this graph are defined on different levels of abstraction (figure 7) and are connected by the common concept of neighborhood that defines the connectivity in the graph.

- If an object is detected but classified as unknown, we can represent it on the lowest level using region information and 2-D spatial relations.
- If an object was classified to a specific object class, we can label the graph node on an intermediate level with this object class.
- If an object has been reconstructed in the 3-D space, we can label the nodes on the higher level with the adapted CAD model, including the features used for the adaptation, such as (contour ellipses and the edges with 3-D spatial relations).

If we label the graph on a higher level, the lower labels still remain because we can use them to get a more probable interpretation of the scene that takes into account failures that happen on a higher-level processing stage.

5 Interaction Scheme

Any verbal description that denotes an object in the scene can be interpreted as a partial representation of this scene (figure 8). If we look at

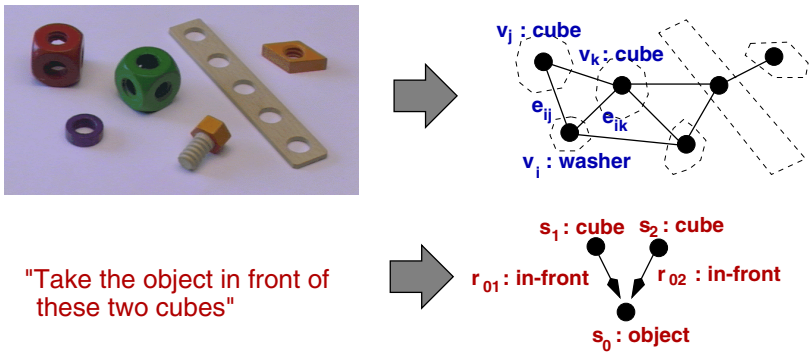


Figure 8. Visual and verbal neighborhood graph of the observed scene.

an utterance, ("Take the object in front of these two cubes"), we know that there are three objects in the neighborhood and that one object lies in front of two cubes. This information can be represented similar to the neighborhood graph that represents the visual information. In order to compute the object that was denoted by the speaker, we have to find the correspondence between the visual and the verbal representation.

5.1 From Verbal Descriptions to Visual Abstraction Hierarchy and Vice-Versa

If a naive user describes an object in the scene by using attributes, he or she will typically use another vocabulary than the fixed vocabulary appropriate for processing visual data without point-to-point correspondence. But, if we want to compare representations generated by vision and by interpretation of verbal descriptions of an object, we need a common concept to connect these two representations: In our domain, this is the object class, which is indented by the speaker and hypothesized by the vision component. Instead of using this concept as a symbolic predicate like classical systems, we model a probability distribution that integrates the abstraction hierarchies and vocabularies on both sides and connects them. Reasoning on this probability distribution is realized by a Bayesian network. A first approach to identifying objects using this network is described in [21].

5.1.1 Verbal descriptions

Verbal descriptions of an intended object class consist of some specified attributes that are partly defined on word level, such as "red," "long," or "bolt," and partly defined on simple grammatical attachments, such as "the bar with three holes." Currently, we distinguish four types of features mentioned in order to specify an intended object class:

- type information (such as bar, bolt, rim, cube, and 3-holed bar),
- color information (such as white, red, dark, and light),
- size information (such as small, big, short, and long), and
- shape information (such as round, hexagonal, and elongated).

All feature types are interpreted as random variables and are modeled as nodes in the Bayesian network. Conditional probabilities $P(\text{feature}_i = f|\text{object\_class} = o)$ connecting these features with an object class are estimated from results of two experiments described in [21]. In the first case, type and color features were extracted from 453 utterances that name a marked object from a scene which was presented on a computer screen. In the second case, size and shape features were collected from a

multiple-choice questionnaire on the Web. All object classes were shown in eight different scene contexts. Participants had to select from the multiple-choice list all features that correctly describe the marked object. A total of 773 people participated in this experiment, and 416 people completed the questionnaire. We estimated the conditional probabilities by counting the uttered type and color features and the selected size and shape features for each object class.

$$P(\text{feature}_i = f | \text{VERBAL\_OBJ\_CLASS} = o) = \frac{\#(\text{feature}_i\ f \text{ is mentioned in the utterance})}{\#(\text{marked object has object class } o)}$$

$$P(\text{feature}_i = f | \text{VERBAL\_OBJ\_CLASS} = o) = \frac{\#(\text{feature}_i\ f \text{ is selected in the questionnaire})}{\#(\text{marked object has object class } o)}$$

5.1.2 Visual abstraction hierarchy

The visual abstraction hierarchy of an object hypothesis consists of several graduated results generated by the vision component. Some components directly hypothesize object classes or subclasses, and others generate results that are related to an object class, such as the color of a region or an ellipse, which is the contour of an object. In our system, we distinguish the following visual evidences:

- HOL_OBJ_CLASS (holistic classification results): object classes or subclasses generated by the holistic component using the special artificial network LLMs.
- REGION_COLOR: color information associated with a segmented region.
- 2D_OBJ_CLASS (2-D classification results): object classes or subclasses, generated by the semantic network.
- CONTOUR_CLASS (contours found in an object region): ellipses, concentric pairs of ellipses, parallel lines, rectangular closed contours, hexagonal closed contours, and so on that are part of the input of the 3-D reconstruction.
- OBJ_CONTOUR_CLASS (contours used in the 3-D reconstruction process): these contours are assigned to features of the geometric object model. Therefore, the possible values of this random variable are features of these geometric models, such as hole-contour of a bar, head-contour of a bolt, body-contour of a bar, and so on.

All types of evidences are modeled as nodes in the Bayesian network. The conditional probabilities are estimated using a labeled test set of 156 objects on eleven images and the recognition results on this test set. Currently, only region color information and 2-D classification results are used in the vision subnet.

$$P(\text{2D\_OBJ\_CLASS} = t | \text{2D\_REG\_OBJ\_CLASS} = o) = \frac{\#(\text{2D\_OBJ\_CLASS } t \text{ was classified as object class } o)}{\#(\text{object has object class } o)}$$

$$P(\text{REGION\_COLOR} = c | \text{2D\_REG\_OBJ\_CLASS} = o) = \frac{\#(\text{REGION\_COLOR } c \text{ was classified as object class } o)}{\#(\text{object has object class } o)}$$

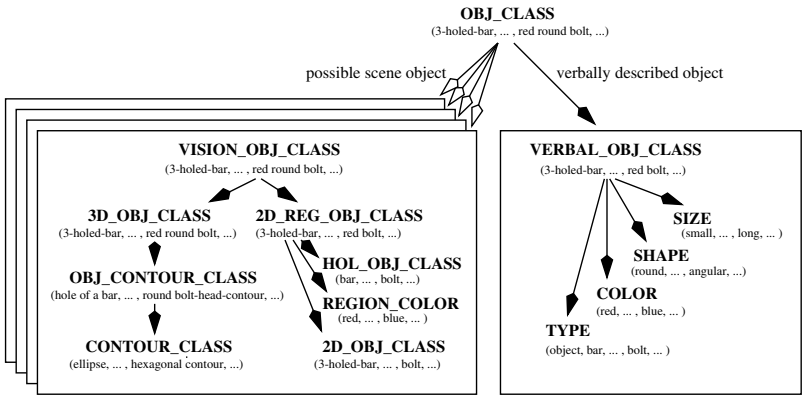


Figure 9. Bayesian network used to connect speech and vision data.

The whole Bayesian network is shown in figure 9. The network provides us with useful information for speech, dialogue, and vision components. If an object, (say an orange screw) is recognized, the 2D_OBJ_CLASS is instantiated with the entry *screw* and the REGION_COLOR with the entry *orange*. Then the Bayesian network can infer that the recognized object has a high probability for the entry *orange_screw* in the VISION_OBJ_CLASS node, a lower probability for a *red_screw*, and a very low probability for a *blue_cube*, because errors with a wrong color classification from *red* to *orange* were found in the recognition results of the test set. On the other side, we can instantiate the evidences from speech understanding in the same manner. For an utterance like "Take the long thin object" the entry *long* can be instantiated in the SIZE node and the entry *thin* in the SHAPE node. Using the Bayesian network, we can infer that a *7-holed bar* is most probably denoted by the utterance, that the *5-holed bar* has a lower probability, and that the *blue_cube* has a very low probability.

5.1.3 Comparison of Verbal Descriptions and Representations from Vision

If evidence has been instantiated in the verbal ($\mathcal{B}_{speech}$) and vision ($\mathcal{B}_{vision}$) subnets of the Bayesian network and a bottom-up propagation is started, we can compare both diagnostic influences on the top object class node in order to measure the correspondence between these two representations. If the two diagnostic influence vectors δ_{speech} and δ_{vision} describe different object classes, they are orthogonal to each other. If they describe the same class, they are in parallel. This is measured by calculating the scalar product of the influence vectors:

$$Cmp(\mathcal{B}_{speech}, \mathcal{B}_{vision}) = \alpha \sum_{i=1}^{\#obj\_classes} \delta_{speech}[i]\, \delta_{vision}[i]$$

where $\delta_{speech}[i]$ is the diagnostic support for an object class i in the subnet $\mathcal{B}_{speech}$,

$\delta_{vision}[i]$ is the diagnostic support for an object class i in the subnet $\mathcal{B}_{vision}$, and

α is a normalizing constant.

After bottom-up and top-down propagation of evidences has been finished and the network is in equilibrium, we get a common belief upon the membership of a scene object to an object class. Two effects can

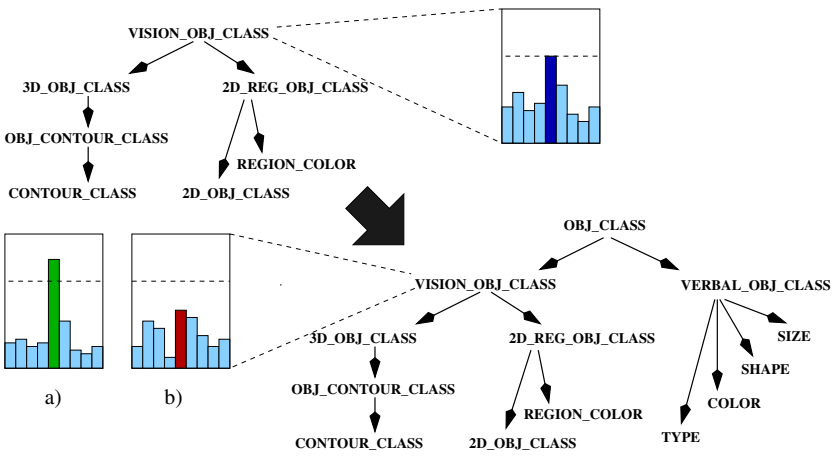


Figure 10. Effects on the belief of an object class considering verbal descriptions.

result from considering the mentioned verbal description: either the gain of information results in an increasing belief in one object class (figure 10a), or the gain of information results in a lower belief in an object class (figure 10b).

If we note the changes of the maximum component of the belief vector the certainty of a node match can be measured.

$$Inf(\mathcal{B}_{speech}, \mathcal{B}_{vision}) = Bel^{vision}_{imax}(\mathcal{E}^{vision}, \mathcal{E}^{speech}) - Bel^{vision}_{imax}(\mathcal{E}^{vision})$$

where $Bel^{vision}_{i}(\mathcal{E})$ is the ith component of the belief vector of the random variable VISION_OBJ_CLASS if the evidence $\mathcal{E}$ is given,

$Bel^{speech}_{i}(\mathcal{E})$ is the ith component of the belief vector of the random variable VERBAL_OBJ_CLASS if the evidence $\mathcal{E}$ is given, and

$imax = \max_i \, Bel^{vision}_{i}(\mathcal{E}^{vision})$

This is important information for the dialogue component that may switch between different dialogue strategies in regards to the certainty measure of the intended object.

- If the selection of the intended object was "certain," the instruction of the user can be directly executed.
- If the selection of the intended object was not "certain," the system can ask for confirmation.
- If the system selects a set of different objects that all had an increasing belief component, the system can ask for a more specific description or otherwise can reject the instruction.

Using the weighted-description comparison and the certainty measure of a selected object, we get a very flexible scoring scheme that integrates evidences from different abstraction levels and offers very detailed information to a dialogue component. The first score measures a kind of "distance" between the visual and the verbal description of an object. The second (the certainty measure) describes a gain of information. The scoring scheme is based on a probability distribution estimated from results of psycholinguistic experiments [21] and from recognition results generated by the vision component.

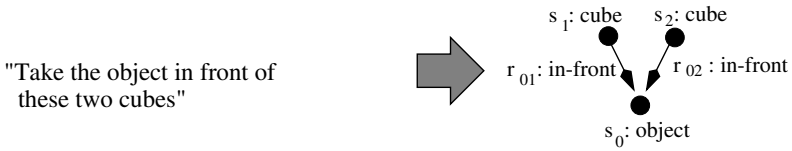


Figure 11. Graph representation generated from an utterance.

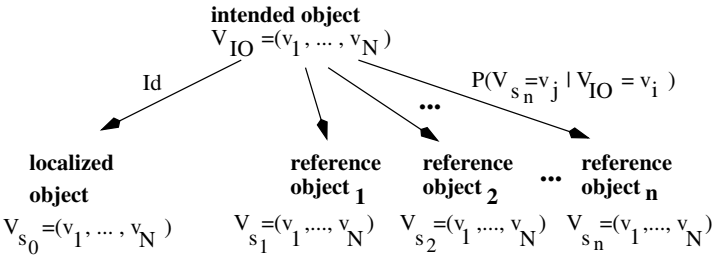


Figure 12. Bayesian network for identifying the intended object.

5.2 Linking Verbal Descriptions and Vision Processing Results by Subgraph Matching

If spatial relations are included in an utterance, the comparison between the verbal description and the visual abstraction hierarchy is expanded to a weighted graph matching problem. Typically, the graph representation of an utterance that denotes an object in the scene is star-shaped (figure 11). It consists of a description of the localized object which is the intended object and a list of reference objects plus associated spatial relations. Several uncertainties have to be considered if we search for the best match of the verbal description and the description of the scene generated by the vision component.

We chose a probabilistic approach, Bayesian networks, to cope with these problems. The structure of the Bayesian network is generated from the graph representation of the verbal description $\mathcal{G}_{speech} = \{\mathcal{S}, \mathcal{R}\}$; the possible labels of the random variables and the conditional probabilities are generated from the graph representation of the vision-processing results $\mathcal{G}_{vision} = \{\mathcal{V}, \mathcal{E}\}$ (figure 12).

$V_s :: \mathcal{V}$ is a random variable that denotes the node of the vision graph which should be assigned to the node s of the verbal graph. The diagnostic support from evidence ϵ^- that the nodes $v \in \mathcal{V}$ and $s \in \mathcal{S}$ are identical is defined as

$$P(\epsilon^-_{V_s=v} | V_s = v) = Cmp(\mathcal{B}_{speech}(s), \mathcal{B}_{vision}(v)),$$

where $\mathcal{B}(v)$ is one of the Bayesian networks $\mathcal{B}_{speech}$ or $\mathcal{B}_{vision}$, which is instantiated with the information attached with a node v.

Another uncertainty that is modeled in the Bayesian network is the probability that $V_{s_k} = v_j$ is the reference object if the user mentions a spatial relation $Rel(r_{0k})$ and we assume that $V_{IO} = v_i$ is the intended object:

$$P(V_{s_k} = v_j | V_{IO} = v_i) = App(Ref, Rel(r_{0\,k}), Rel(e_{ij}))$$

where $Rel(e_{ij}) = Rel(Obj(v_i), Obj(v_j))$ is the representation of the spatial relation between the objects $Obj(v_i)$ (IO) and $Obj(v_j)$ (RO),
Ref is the current reference frame of the speaker, and
$App(\ldots)$ denotes the applicability function of the spatial model.

If we want to propagate the evidences to the top node, we have to consider some constraints. One particular node $v \in \mathcal{V} = \{v_1 \ldots v_N\}$ can be

assigned to only one of the nodes $s \in \mathcal{S} = \{s_0 \ldots s_n\}$, and all assignments must be injective. This concerns a search algorithm that finds the most plausible assignment of nodes for all values of $V_{IO} = v_{io=1\ldots N} \in \mathcal{V}$, which is defined by

$$
\begin{aligned}
&(v_{s_0} \ldots v_{s_n})^* \\
&\quad = \underset{(v_{s_0} \ldots v_{s_n})}{argmax}\ Bel(V_{IO} = v_{io}) \\
&\quad = \underset{(v_{s_0} \ldots v_{s_n})}{argmax}\ P(\epsilon^{-}_{V_{s_0}=v_{s_0}} \ldots \epsilon^{-}_{V_{s_n}=v_{s_n}} | V_{IO} = v_{io}) \\
&\quad = \underset{(v_{s_0} \ldots v_{s_n})}{argmax} \prod_{i=0\ldots n} P(\epsilon^{-}_{V_{s_i}=v_{s_i}} | V_{s_i} = v_{s_i}) P(V_{s_i} = v_{s_i} | V_{IO} = v_{io})
\end{aligned}
$$

where $(v_{s_0} \ldots v_{s_n}) \in \mathcal{V}^n$ are injective node assignments for $(s_0 \ldots s_n)$. Propagating down gives us some expectations about an object that would be consistent with the rest of our instantiated evidence.

Evaluating the Bayesian network provides a plausibility for all objects in the scene. The components of this plausibility vector rate what object was denoted by the spoken utterance. If we have to chose a particular object to generate a system answer, many times the decision cannot be definite. This will happen for several reasons.

- The user intended to specify a group of objects.
- The user did not realize that the naming was not specific enough.
- Some attributes the user mentioned were not recognized by speech processing.
- The vision component classified an object to the wrong class.

The ambiguity remaining in the set of selected objects can be resolved only by a dialog component that is able to consider the dialog context of the utterance. Therefore, the answer of the component that interprets the utterance only in scene context has to include all objects that are plausible.

Partitioning the plausibility vector U_{IO} into one group of elements that defines the system answer and one group with plausibility below a threshold is a difficult task because we have no information about a distribution of these values on which such a decision could be based. Often, we intuitively split such vectors into more than two groups, (for example, one with low, one with middle, and one with high plausibility).

Therefore, we use differences of values in order to partition the plausibility vector (figure 13). In a first step, the plausibility values are sorted (1.) so that we get a monotonic discrete-sized function. The biggest jumps in this function are hypothetical boundaries of partitions. In order to get the right number of partitions, the differences of each jump are sorted (2.+3.). The maximum jump in this new monotonic function (4.) provides all the differences that are significant for partitioning the original sorted plausibility vector (5.+6.+7.). Finally, the system answer is generated, including all objects that are member of the partition with maximal plausibility (8.).

The partitioning scheme can be influenced by varying the zero-partitioning line. This value defines the difference that is assigned to the minimum component of the plausibility vector. If it is set to zero, there is a tendency to select all objects in context. If it is set to the value of the minimum component, there are more differences that are classified as significant (figure 14). This may be useful if the scene context is

Figure 13. Example presenting the partitioning of the plausibility vector.

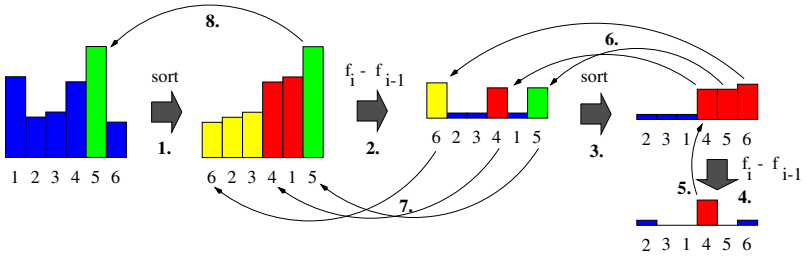
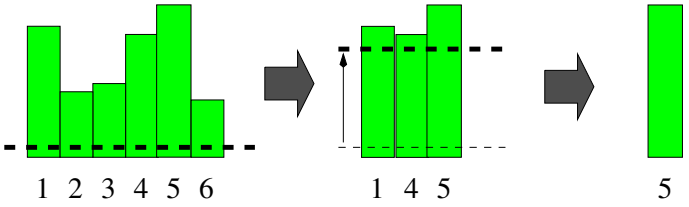


Figure 14. Increasing the zero-partitioning line in different focus contexts. If the maximum component of the example in figure 13 is slightly less, three components are selected. If we focus on these and raise the zero-partitioning line, the maximum component can be selected.



varied by the dialogue component relative to a set of objects that may be in focus.

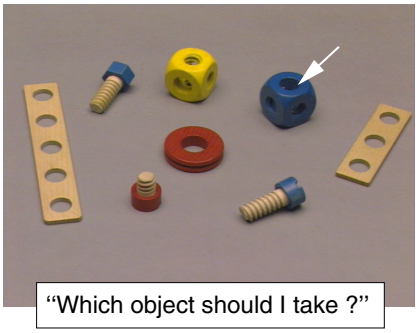
6 Results

In this section, we will concentrate on aspects concerning the interaction scheme of the speech-understanding and vision component. Given an instruction, the system has to identify the objects denoted by the utterance. The input of the system is only one utterance and only one image. Therefore, it is a quite restricted evaluation because in the assembling scenario is a dialogue between the instructor and the system, and the object recognition uses an image sequence in order to stabilize the vision results. The results presented here should be interpreted in this context.

First, we mention some aspects concerning the results of the system components. Second, we compare the identification results of our system under different conditions using speech and transcribed (NL) input, uncertain and perfect vision. In this section, we present the main aspects of our first results and give some interpretations of them. (You can get a more detailed description of the results in the Web document.)

6.1 Experimental Data

Two different test sets were used for evaluating our system. Figure 15 gives an impression of the experiment we used to collect both sets of spoken utterances. In the first psycholinguistic experiment ten subjects verbally named objects that were presented on a computer screen. In each image, one object was marked and the subjects were told to name this object using an instruction, such as “take the big blue object” or “take the long bar.” From this experiment, 453 verbal object descriptions were collected. These descriptions were partly used to estimate some conditional probabilities concerning type and color information in the Bayesian network used for node comparison (section 5.1). This test set is called FEATURE set. In a second experiment under the same conditions, six subjects were asked to name the marked object by using spatial relations, such as “take the blue one behind the screw.” From this, 144 utterances were collected using five different scenes. This test set is called SPATIAL set.



test-set 1 (FEATURE) : "Take the big blue Object."
test-set 2 (SPATIAL) : " Take the blue one behind the screw."

Figure 15. Example of an image used for the evaluation experiments. The person has to name the marked object by specifying some features of the object or using spatial relations.

Table 1. Results of the speech components on whole test sets.

	#utt	WA	FA
FEATURE_S+SPATIAL_S	448	66.6%	77.4%

Table 2. Results of the vision component on different test sets.

	#utt	PV	**color error**	**object class error**	**none**
FEATURE_S	325	88.9%	5.2%	2.2%	3.7%
SPATIAL_S	123	71.5%	20.3%	4.1%	4.1%

The instructions were transcribed orthographically to exclude speech-recognition errors. This input is labeled as NL (natural language). Instead of a set of labeled images, we use a subset of utterances that denote a correctly recognized object. This input is labeled as PV (perfect vision).

6.2 Component Results

In order to interpret the identification results correctly, we have to analyze the partial results of the system components. On the speech side, the quality of the recognition component is measured by the word accuracy (WA) [11]. A similar measure can be defined for feature structures, counting substitutions, insertions, and deletions of feature entries. For that, we use the feature structures as reference data that are generated by the speech-understanding component using NL input. This measure is called feature accuracy (FA) and can be interpreted as the influence of speech-recognition errors on the speech-understanding component for object specifications. Even for NL input, some failures occur in the understanding component, due to unknown words or language structures that were not modeled in the grammar or in the semantic network. In these cases the understanding component generated none or more than one object specification. These utterances were disregarded when computing the identification results. The word and feature accuracy for both speech test sets is given in table 1.

On the vision side are three different kinds of errors: wrong object class, wrong color classification, and no object hypothesis at all. In some scenes, an object was unknown to the vision component. These objects caused most wrong-object classifications. We measured these failures for all marked objects of the different test sets (table 2).

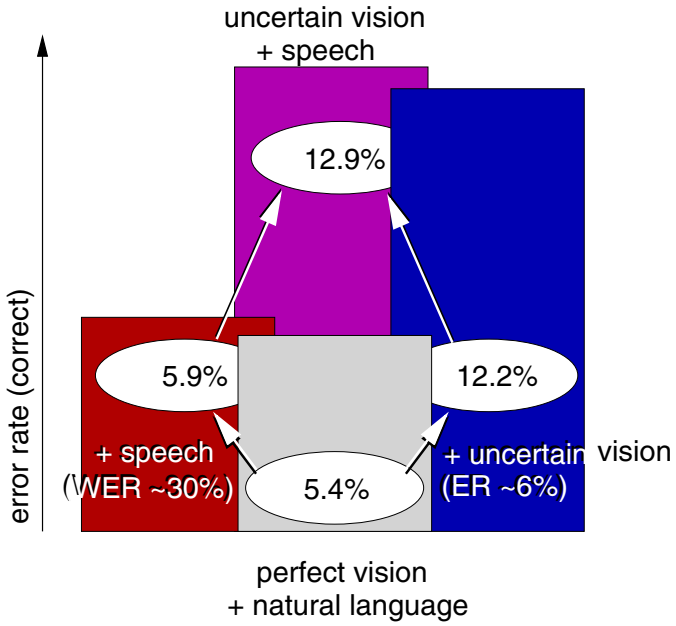


Figure 16. Correct system answers for the FEATURE set using NL/speech and perfect/uncertain vision.

6.3 Classification of System Answers

Given an instruction, the system generates a set of objects that were hypothetically denoted by the instruction. The system answer is classified into different classes in regard to the intended marked object.

- **precise:** the marked object is the only one the system selected
- **included:** the system selected more objects beside the marked one which have the same object class as the marked object
- **additional:** the marked object is member of the selected subset of objects, but some selected objects have a different object class than the marked object
- **correct:** the marked object is a member of the selected subset. Note that "correct" is the union of "precise," "included," and "additional"
- **false:** the system has selected some objects, but the marked object is not a member of the subset
- **nothing:** the system rejected the instruction because the system did not find an appropriate object.

The precise class is relevant only if the intended object is localized by specifying some reference objects and spatial relations. Otherwise, the system cannot distinguish objects of the same object class.

In the following subsections, we concentrate on the error rate of correct system answers. If the intended object is not selected, the dialogue component doesn't have enough information to select a goal-directed stategy. If the system selects more objects than the intended object, there is always the chance to get the intended object in the next dialogue step simply by asking the right question. So this measure is very sensitive for the acceptance of the system to the user.

6.4 Results Using the FEATURE Set

In figure 16, we present the identification results using the FEATURE set of instructions. The base error rate (PV+NL) reflects some aspects every natural human-computer interface:

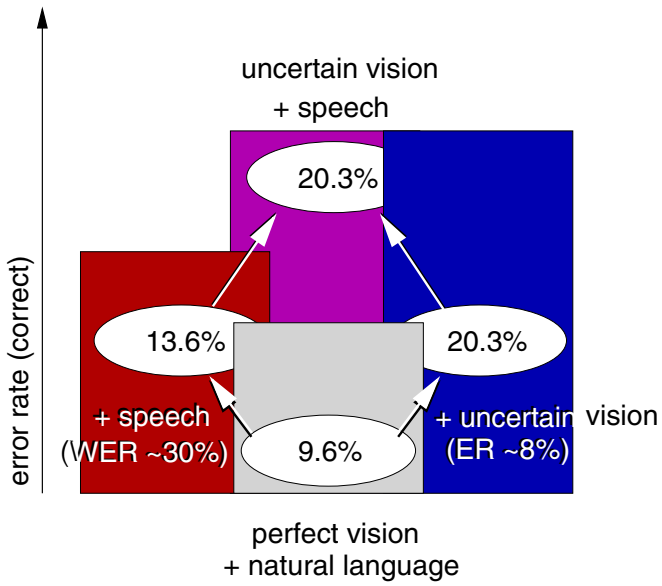


Figure 17. Correct system answers for the SPATIAL set using NL/speech and perfect/uncertain vision.

- The instructor used some attributes or statements that were not modeled in the system.
- The instructor did not specify the marked object precisely enough. Even a human sometimes cannot figure out the intended object.

When we introduce uncertainties, it is remarkable that the impact of speech input is negligible even though the word accuracy (66.6%) was quite worse and the feature accuracy (77.4%) is far from that of text input. The rate of system answers classified as additional was below 9% in all cases. Therefore, the tendency to select too many objects is quite low. The additional rate rose from 6.5% to 9% when introducing speech input.

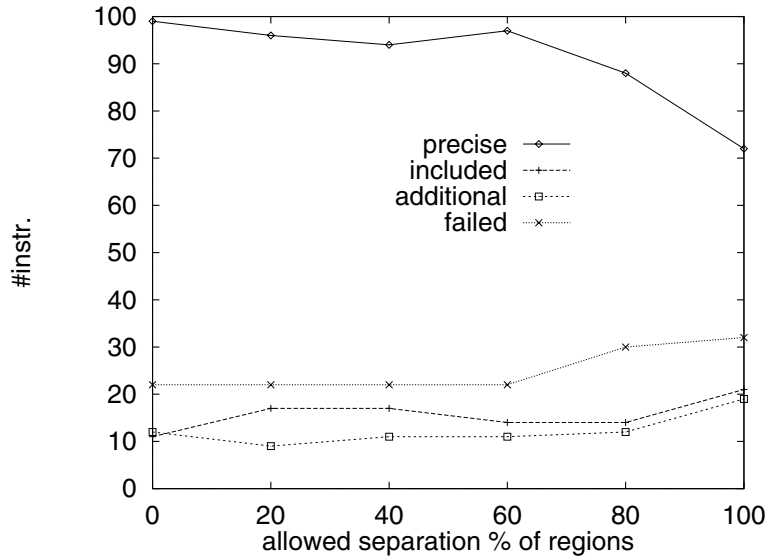
The error rate of the vision component is nearly added to the error rate of correct system answers. Especially when the object is totally lost or the classified object type has changed, there is no (or only a little) change to get the right object. Whether the system can select the right object in such a case is fundamentally dependent on the redundancy in the verbal object specification.

6.5 Results Using the SPATIAL Set

In the previous section, we presented identification results from utterances that did not mention spatial relations. Therefore, only the Bayesian networks (for verbal feature descriptions and the visual abstraction hierarchy) and the partitioning scheme of the plausibility vector are evaluated. In the SPATIAL set, we have to consider more influences on the identification results, (for example spatial model and topology). The results are presented in figure 17. In this graph, we can examine the same tendencies as mentioned for the FEATURE set. The base error rate is higher because there are more variations in the utterances and the identification of the intended object depends highly on the identification of the reference object. Another consequence of this dependency is that the error rate of the vision component is very sensitive for the identification results. On the other side, the influence of speech errors is very low.

In order to measure the influence of topology introduced by the neighborhood graph on the system answers, we analyzed the rate of precise

Figure 18. Influence of the neighborhood graph.



system answers on the SPATIAL test set using NL/UV input. The neighborhood of objects in the scene is defined by object separations (section 4). We varied the threshold defining the percentage by which two object regions are allowed to be separated without losing their neighborhood. (A value of 100% means that all objects are neighbors.) The classification of the system answers are shown in figure 18. We observe that above 60% separation, system answers are much more sloppy and that additional and false system answers increase. On the other side, we nearly did not lose any precise system answer if the threshold is defined very rigorously. This provides good evidence that the assumptions on which we defined the neighborhood hold. The concept of defining neighborhood relations by separations seems to model an important aspect when humans specify intended objects using spatial relations.

7 Conclusion

We presented a new approach to integrating vision and speech understanding. As we work in real environments, we are typically concerned with erroneous, vague, or incomplete data. We cope with these problems by using multiple representations on different levels of abstraction. In order to find the most plausible interpretation of the scene, the graduated results of the vision components are connected by Bayesian networks.

The interpretation of a spoken utterance concerning the visual scene is modeled as another partial representation of the scene. Verbal descriptions often use qualitative features of the object instead of the exact name and specify spatial relations only vaguely. We model the coherence between these features and the object classes used in vision in a Bayesian network. The conditional probabilities were estimated using results of psycholinguistic experiments.

We showed that the topology of the scene is an important aspect if a speaker localizes an intended object by using spatial relations. We introduced this influence by the definition of a neighborhood graph that is used on all representation layers. The nodes of this graph are labeled with the graduated results generated by the vision components respectively with mentioned features of a verbal object description. The edges

are labeled with spatial relations which are represented numerically (vision) or by name (verbal).

The identification of the object that was indented by the speaker is modeled as a weighted subgraph match between the graph representation generated by vision and the verbally specified graph representation. The weighting scheme is realized by a Bayesian network. For each object in the scene, the plausability is calculated that it was denoted by the utterance. In order to select a set of objects that are hypothetically denoted by the utterance, we developed a partitioning technique which does not need a prespecified threshold and does not suppose a specific distribution of the values. Secondly, we defined an additional confidence measure that provides a dialogue component some information concerning the uncertainty of the selection.

We demonstrated the effectiveness of our approach on real data in which naive users were asked to name objects in a visual scene. We could make the spatial inferences in our system more precise by the introduction of the neighborhood graph and more robust by modeling them on different abstraction layers. Because of the hierarchical modeling, we obtain first results with drastically reduced computational complexity. As computation continues, results will be more precise. Thereby, an anytime behavior is realized.

Further research will concentrate on the expansion of the vision abstract hierarchy, the modeling of an increased set of spatial relations, and the integration of hand gestures pointing at an intended object. We will also use the plausibility measure from object identifications as a scoring scheme in the understanding component to weight alternative interpretations of the utterance.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) in the project SFB 360 “Situated Artificial Communicators.”

References

[1] G. Adorni, M. Di Manzo, and F. Giunchiglia. Natural language driven image generation. In *COLING*, pages 495–500, 1984.

[2] L. E. Bernstein. For speech perception by humans or machines, three senses are better than one. In *International Conference on Spoken Language Processing*, pages 1477–1480, 1996.

[3] H. Brandt-Pook, G. A. Fink, S. Wachsmuth, and G. Sagerer. Integrated recognition and interpretation of speech for a construction task domain. In *Proceedings of the International Conference on Human Computer Interaction (HCI)*, 1, pages 550–554, 1999.

[4] S. Dickenson and D. Metaxas. Integrating qualitative and quantitative shape recovery. *International Journal of Computer Vision*, 13(3):1–20, 1994.

[5] T. Fuhr, G. Socher, C. Scheering, and G. Sagerer. A three-dimensional spatial model for the interpretation of image data. In P. Olivier and K.-P. Gapp (Eds.), *Representation and Processing of Spatial Expressions*, pages 103–118. Lawrence Erlbaum Associates, 1997.

[6] G. Heidemann, H. Ritter, F. Kummert, and G. Sagerer. A Hybrid Object Recognition Architecture. In C. von der Malsburg, W. von

Seelen, J. C. Vorbruggen, and B. Sendhoff (Eds.), *Artificial Neural Networks*, pages 305–310. Springer-Verlag, Berlin, 1996.

[7] H. Kollnig and H.-H. Nagel. Ermittlung von begrifflichen Beschreibungen von Geschehen in Straßenverkehrsszenen mit Hilfe unscharfer Mengen. In *Informatik Froschung und Entwicklung*, 8, pages 186–196, 1993.

[8] S. M. Kosslyn. Mental imagery. In D. A. Osherson et al. (Eds.), *Visual Cognition and Action*, pages 73–97. MIT Press, Cambridge, Mass, 1990.

[9] F. Kummert, G. A. Fink, G. Sagerer, and E. Braun. Hybrid Object Recognition in Image Segvenus. 14th International Conference On Pattern Recognition, volume II, pages 1165–1170. Brisbane, 1998.

[10] F. Lavagetto, S. Lepsoy, C. Braccini, and S. Curinga. Lip motion modeling and speech driven estimation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 183–186, 1997.

[11] K. F. Lee. *Automatic Speech Recognition: The Development of the SPHINX System*. Kluwer Academic Publishers, 1989.

[12] A. Maßmann, S. Posch, and D. Schlüter. Using Markov random fields for contour-based grouping. In *Proceedings of International Conference on Image Processing*, volume 2, pages 207–242, 1997.

[13] T. Maybury (Ed.). *Intelligent Multimedia Interfaces*. AAAI Press/The MIT Press, 1993.

[14] D. McDonald and E. J. Conklin. Salience as a simplifying metaphor for natural language generation. In *Proceedings of AAAI-81*, pages 49–51, 1981.

[15] K. Nagao. Abduction and dynamic preference in plan-based dialogue understanding. In *International Joint Conference on Artificial Intelligence*, pages 1186–1192. Morgan Kaufmann Publishers, Inc., 1993.

[16] K. Nagao and Jun Rekimoto. Ubiquitous talker: Spoken language interaction with real world objects. In *International Joint Conference on Artificial Intelligence*, pages 1284–1290, 1995.

[17] P. Olivier, T. Maeda, and J. Ichi Tsujii. Automatic depiction of spatial descriptions. In *Proceedings of AAAI-94*, pages 1405–1410, Seattle, WA, 1994.

[18] W. Richards, A. Jepson, and J. Feldman. Priors, preferences and categorial percepts. In W. Richards and D. Knill (Eds.), *Perception as Bayesian Inference*, pages 93–122. Cambridge University Press, 1996.

[19] G. Socher, T. Merz, and S. Posch. 3-D reconstruction and camera calibration from images with known objects. In D. Pycock, (Ed.), *Proceedings Sixth British Machine Vision Conference*, pages 167–176, 1995.

[20] G. Socher. *Qualitative Scene Descriptions from Images for Integrated Speech and Image Understanding*. Dissertationen zur Künstlichen Intelligenz (DISKI 170). Infix-Verlag, Sankt Augustin, 1997.

[21] G. Socher, G. Sagerer, and P. Perona. Bayesian reasoning on qualitative descriptions from images and speech. In H. Buxton and

A. Mukerjee (Eds.), *ICCV'98 Workshop on Conceptual Description of Images*, Bombay, India, 1998.

[22] R. K. Srihari. Computational models for integrating linguistic and visual information: A survey. In *Artificial Intelligence Review*, 8, pages 349–369. Kluwer Academic Publishers, Netherlands, 1994.

[23] R. K. Srihari and D. T. Burhans. Visual semantics: Extracting visual information from text accompanying pictures. In *Proceedings of AAAI-94*, pages 793–798, Seattle, WA, 1994.

[24] J. K. Tsotsos et al. The PLAYBOT Project. In J. Aronis (Ed.), *IJCAI '95 Workshop on AI Applications for Disabled People*, Montreal, 1995.

[25] J. K. Tsotsos, G. Verghese, S. Dickenson, M. Jenkin, A. Jepson, E. Milios, F. Nuflo, S. Stevenson, M. Black, D. Metaxas, S. Culhane, Y. Yet, and R. Mann. PLAYBOT: A visually-guided robot for physically disabled children. *Image and Vision Computing*, 16(4):275–292, 1998.

[26] G. Verghese and J. K. Tsotsos. Real-time model-based tracking using perspective alignment. In *Proceedings of Vision Interface '94*, pages 202–209, 1994.

[27] C. Vorwerg, G. Socher, T. Fuhr, G. Sagerer, and G. Rickheit. Projective relations for 3-D space: computational model, application, and psychological evaluation. In *Proceedings of the 14th National Joint Conference on Artificial Intelligence AAAI-97*, Rhode Island, 1997.

[28] S. Wachsmuth, G. A. Fink, and G. Sagerer. Integration of parsing and incremental speech recognition. In *Proceedings EUSIPCO-98*, 1998.

[29] W. Wahlster. One word says more than a thousand pictures: On the automatic verbalization of the results of image sequence analysis systems. In *Computers and Artificial Intelligence*, 8, pages 479–492, 1989.

[30] D. L. Waltz. Generating and understanding scene descriptions. In Bonny Webber and Ivan Sag (Eds.), *Elements of Discourse Understanding*, pages 266–282. Cambridge University Press, New York, 1981.

[31] M. Zancanaro, O. Stock, and C. Strapparava. Dialogue cohension sharing and adjusting in an enhanced multimodal environment. In *International Joint Conference on Artificial Intelligence*, pages 1230–1236. Morgan Kaufmann Publishers, Inc., 1993.