Bayesian Networks for Speech and Image Integration

Sven Wachsmuth and **Gerhard Sagerer**
Bielefeld University, Faculty of Technology, 33594 Bielefeld, Germany
swachsmu@techfak.uni-bielefeld.de

Abstract

The realization of natural human-computer interfaces suffers from a wide range of restrictions concerning noisy data, vague meanings, and context dependence. An essential aspect of everyday communication is the ability of humans to ground verbal interpretations in visual perception. Thus, the system has to be able to solve the correspondence problem of relating verbal and visual descriptions of the same object. This contribution proposes a new and innovative solution to this problem using Bayesian networks. In order to capture vague meanings of adjectives used by the speaker, psycholinguistic experiments are evaluated. Object recognition errors are taken into account by conditional probabilities estimated on test sets. The Bayesian network is dynamically built up from verbal object description and is evaluated by an inference technique combining bucket elimination and conditioning. Results show that speech and image data is interpreted more robustly in the combined case than in the case of isolated interpretations.

Introduction

Speech understanding and vision are the most important abilities in human-human communication, but also the most complex tasks for a machine. Typically, speech understanding and vision systems are realized for a dedicated application in a constrained domain. Both tasks are realized using different specialized paradigms and separated knowledge bases. They use different vocabularies to express the semantic content of an input signal. Consequently, the *correspondence problem* – namely how to correlate visual information with words, events, phrases, or entire sentences – is not easy to solve (Srihari 1994). A human speaker encodes the verbal-visual correspondences in an internal representation of the sentence he or she intends to utter. The communication partner has to decode these correspondences without knowing the mental models and internal representation of the speaker. Thus, *referential uncertainty* is inherently introduced even for perfect understanding components. Additionally, the interpretations of the signal modalities are often erroneous or incomplete such that an integrating component must consider noisy and partial interpretations. As a consequence, this contribution treats the correspondence problem as probabilistic *decoding process*. This perspective distinguishes this approach from other approaches that propose rule-based translation schemes (Takahashi *et al.* 1998) or integrated knowledge bases (Brondsted *et al.* 1998; Srihari & Burhans 1994). These assume that a visual representation can be logically transformed into a verbal representation and vice versa.

An important issue for a system that relates erroneous and incomplete interpretations is *robustness*, i.e. how the system answer is affected by propagated errors. The proposed approach shows that even though a multi-modal system has to face multiple error sources *the combined signal can be interpreted more stably than the individual signals*. This is explicitly shown by a detailed analysis of the identification rates of an implemented system in a construction domain.

It reveals that the Bayesian network used for integration of speech and image interpretations has to combine spatial and type information. Instead of modeling these two kinds of evidence in separated Bayesian networks (Rimey & Brown 1994; Intille 1999), a novel Bayesian network scheme is developed by introducing *selection variables* and exploiting the properties of these variables during inference.

Object descriptions from vision and speech

Interpretations extracted from speech and images can only be integrated on the basis of a common notion, e.g. a visually perceived and verbally mentioned object (cf. (Jackendoff 1987)). On the speech side an object instance is verbally

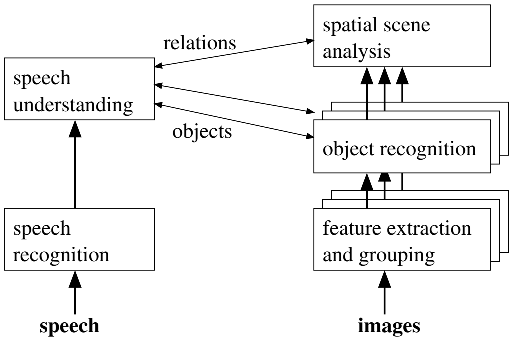


Figure 1: Object descriptions extracted from speech and images

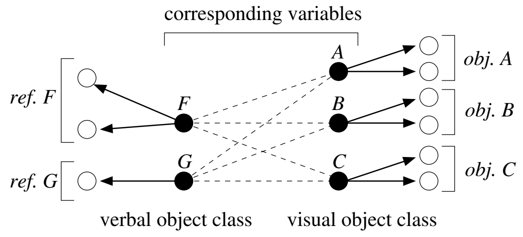


Figure 2: The correspondence problem in Bayesian networks: Two objects F,G referred by the speaker has to be related to three visually detected objects A,B,C.

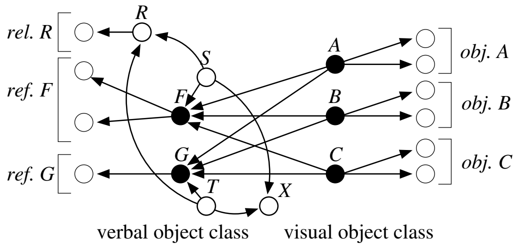


Figure 3: Modeling of the correspondence problem using selection variables S,T.

described by nouns, adjectives, and spatial relations to other object instances. On the vision side, an object instance is described by the location of grouped features, feature values, and the label provided by object recognition algorithms. Additionally, a qualitative scene description may be derived from the object locations defining relations between objects (Fig. 1). Both processes are error-prone and are typically based on separated knowledge bases. Thus, the task of establishing correspondences between vision and speech processing results should consider both processes as well as the mapping between them as probabilistic.

Bayesian networks

Bayesian networks (Pearl 1988) model a joint probability distribution over a set of random variables $\mathcal{U} = \{A_1,\ldots,A_n\}$ with discrete states. Conditional independencies between sets of variables are represented in a directed acyclic graph $\mathcal{G}$.

$$P(\mathcal{U}) = \prod_{A\in\mathcal{U}} P(A|pa(A))$$

where $pa(A)$ denotes all parents of node A in $\mathcal{G}$

In the following, upper letters denote random variables, lower letters denote states of variables. Thus, $P(A|B,C)$ is a conditional probability table (CPT) while

$$P(a_i|b_j,c) = P(A=a_i|B=b_j,C=c)$$

is the probability value after assigning the i-th state of A, the j-th state of B, and state c. Sets of variables are distinguished by raised indices, e.g. $A^{(i)} \in \{a_1^{(i)},\ldots,a_n^{(i)}\}$.

Modeling of the correspondence problem

The correspondence problem is formulated by several subnetworks that include a set of *corresponding variables* (see Fig. 2), e.g. variables modeling the object classes for the sentence *“Take the small ring in front of the rotor”* and the scene in Fig. 6. Each variable on the speech side corresponds to another variable on the vision side, but the correct assignment is not previously known. It has to be inferred from evidences. Thus, the correspondence property has to be modeled in the language of Bayesian networks. Let F be a speech variable that corresponds to one of the vision variables A or B. This *one-to-two* mapping is represented by a *selection variable* $S \in \{\tilde{a},\tilde{b}\}$ and the conditional probability table

$$P(F|A,B,S) = [P(f_i|a_j,b_k,s); i=1\ldots m, j=1\ldots n, \\ k=1\ldots r, s\in\{\tilde{a},\tilde{b}\}]$$

$$\text{where } P(f_i|a_j,b_k,s) = \begin{cases} P(f_i|a_j) & \text{, if } s=\tilde{a} \\ P(f_i|b_k) & \text{, if } s=\tilde{c} \end{cases}$$

If $S=\tilde{a}$, B is irrelevant to F. If $S=\tilde{b}$, A is irrelevant to F. A *one-to-N* mapping between variable F and variables $A^{(1)}\ldots A^{(N)}$ can be modeled by extending the possible states of $S\in\{\tilde{a}_1,\ldots,\tilde{a}_N\}$:

$$P(f|a^{(1)},\ldots,a^{(N)},s) = P(f|a^{(i)}), \text{if } s=\tilde{a}_i$$

Exclusive *M-to-N* mappings between variables $F^{(1)}\ldots F^{(M)}$ and $A^{(1)}\ldots A^{(N)}$ are modeled by introducing M selection variables with N states each, $S^{(1)},\ldots,S^{(M)}\in\{\tilde{a}_1,\ldots,\tilde{a}_N\}$:

$$P(f^{(i)}|a^{(1)},\ldots,a^{(N)},s^{(i)}) = P(f^{(i)}|a^{(j)}), \text{if } s^{(i)}=\tilde{a}_j \\ \text{where } 1\leq i\leq M, 1\leq j\leq N$$

and an exclusive relation R with

$$P(X=1|s^{(1)},\ldots,s^{(M)}) \\ = \begin{cases} 1.0, & \text{if } s^{(i)}\neq s^{(k)}, 1\leq i,k\leq M, i\neq k \\ 0.0, & \text{otherwise} \end{cases}$$

Spatial or structural information about verbally mentioned objects can be coded by constraining possible values of selection variables. These constraints are extracted from visual information that is represented in $P(R|S,T)$ (see subsection about **Spatial modeling**). The resulting network is presented in Fig. 3.

Inference in Bayesian networks

The solution of the correspondence problem in a Bayesian network including M selection variables is a *maximum a posteriori hypothesis* (map) task. We are searching for the most probable states of the selection variables $S_1,\ldots,S_M$ with regard to the marginal distribution arising out of the observed evidences $\mathbf{e}_F^-,\mathbf{e}_G^-,\mathbf{e}_A^-,\mathbf{e}_B^-,\mathbf{e}_C^-,\mathbf{e}_R^-$. The inference algorithm used in the system presented in this paper is based on the *bucket elimination* algorithm proposed by Rina Dechter

(Dechter 1998). It can be formulated in the algebra of conditional probability tables (CPTs) as a summation and maximization over a product of CPTs:

$$\begin{aligned}s^*,t^* = & \underset{s,t}{\operatorname{argmax}}\ P(X=1|S,T)\\ & \sum_R P(\mathbf{e}_R^-|R)\ P(R|S,T)\ \sum_A P(\mathbf{e}_A^-|A)\ P(A)\\ & \sum_B P(\mathbf{e}_B^-|B)\ P(B)\ \sum_C P(\mathbf{e}_C^-|C)\ P(C)\ \sum_F P(\mathbf{e}_F^-|F)\\ & \quad P(F|A,B,C,S)\ \sum_G P(\mathbf{e}_G^-|G)\ P(G|A,B,C,T)\end{aligned}$$

where each summation or maximization over a variable defines a bucket. Thus, the result of a summation can be interpreted as a message from one bucket to another. A straight forward evaluation of a Bayesian network that includes M selection variables with N states each results in messages to be propagated which sizes are exponential in M. However, The CPTs $P(F|A,B,C,S)$ and $P(G|A,B,C,T)$ define a context-specific independence (CSI) as discussed in (Boutilier *et al.* 1996). Instead of introducing numerous additional nodes to the Bayesian network in order to cover the CSI in the structure of the network, here, we propose an algorithmic solution by combining conditioning and bucket elimination techniques. This approach has the advantage of being better suited for extension towards approximate inference. Using conditioning over the selection variables S,T, the inference algorithm can be formulated as a conditioned sum:

$$\begin{aligned}s^*,t^* = & \underset{s,t}{\operatorname{argmax}}\ P(X=1|S,T)\\ & \sum_R P(\mathbf{e}_R^-|R)\ P(R|S,T)\ \sum_A P(\mathbf{e}_A^-|A)\ P(A)\\ & \sum_B P(\mathbf{e}_B^-|B)\ P(B)\ \sum_C P(\mathbf{e}_C^-|C)\ P(C)\\ & \sum_F P(\mathbf{e}_F^-|F)\begin{cases}P(F|A,S=\tilde{a})\\ P(F|B,S=\tilde{b})\\ P(F|C,S=\tilde{c})\end{cases}\\ & \sum_G P(\mathbf{e}_G^-|G)\begin{cases}P(G|A,T=\tilde{a})\\ P(G|B,T=\tilde{b})\\ P(G|C,T=\tilde{c})\end{cases}\end{aligned}$$

After evaluation of the last two buckets, i.e. summations over G,F, and propagation of the resulting messages $\lambda_F(\dots)$ and $\lambda_G(\dots)$ we get:

$$\begin{aligned}s^*,t^* = & \underset{s,t}{\operatorname{argmax}}\ P(X=1|S,T)\\ & \sum_R P(\mathbf{e}_R^-|R)\ P(R|S,T)\\ & \sum_A P(\mathbf{e}_A^-|A)\ P(A)\begin{cases}\lambda_F(A,S=\tilde{a})\begin{cases}\lambda_G(A,T=\tilde{a})\\ \lambda_G(A,T\neq\tilde{a})\end{cases}\\ \lambda_F(A,S\neq\tilde{a})\begin{cases}\lambda_G(A,T=\tilde{a})\\ \lambda_G(A,T\neq\tilde{a})\end{cases}\end{cases}\\ & \sum_B P(\mathbf{e}_B^-|B)\ P(B)\begin{cases}\lambda_F(B,S=\tilde{b})\begin{cases}\dots\end{cases}\\ \lambda_F(B,S\neq\tilde{b})\begin{cases}\dots\end{cases}\end{cases}\\ & \sum_C P(\mathbf{e}_C^-|C)\ P(C)\begin{cases}\dots\end{cases}\end{aligned}$$

Evaluating buckets A,B,C,R, then, yields:

$$s^*,t^* = \underset{s,t}{\operatorname{argmax}}\ P(X=1|S,T)\ \lambda_R(S,T)\ \lambda_{A,B,C}(S,T)$$

where $\lambda_{A,B,C}(S,T)$

$$\begin{aligned}= & \begin{bmatrix}\lambda_A(S=\tilde{a},T=\tilde{a}) & \lambda_A(S\neq\tilde{a},T=\tilde{a})\\ \lambda_A(S\neq\tilde{a},T=\tilde{a}) & \lambda_A(S\neq\tilde{a},T\neq\tilde{a})\end{bmatrix}\\ & \cdot\begin{bmatrix}\lambda_B(S=\tilde{b},T=\tilde{b}) & \lambda_B(S\neq\tilde{b},T=\tilde{b})\\ \lambda_B(S\neq\tilde{b},T=\tilde{b}) & \lambda_B(S\neq\tilde{b},T\neq\tilde{b})\end{bmatrix}\\ & \cdot\begin{bmatrix}\lambda_C(S=\tilde{c},T=\tilde{c}) & \lambda_C(S\neq\tilde{c},T=\tilde{c})\\ \lambda_C(S\neq\tilde{c},T=\tilde{c}) & \lambda_C(S\neq\tilde{c},T\neq\tilde{c})\end{bmatrix}\end{aligned}$$

In general, the evaluation scheme still holds if the different corresponding sub-networks are not d-separated by the variable A,B,C,F, or G, respectively. In this case the conditioning structure of each bucket is more complex and the final message $\lambda_{A,B,C}$ cannot be defined by three independent factors. In practice, this causes no problem as long as the problem size is small or the subnets can be d-separated by instantiating evidential variables. As an additional step, the properties of the CPT $P(X=1|S,T)$ are exploited during evaluation such that not all conditional summations have to be computed.

Application to a construction scenario

The Bayesian network approach for integrating speech and images has been applied to a *situated artificial communicator* in a construction scenario. The project aims at the development of a robot constructor which can be instructed by speech and gestures in the most natural way (Bauckhage *et al.* 2001).

While the communication between the human instructor and the system is constrained as little as possible, the domain setting is rather restricted. A partial collection of 23 different elementary components of a wooden toy construction kit is placed on a table. The elementary objects have characteristic colors. However, the colors do not uniquely determine the class of an object. The table scene is perceived by a calibrated stereo camera head that is used for object recognition and localization. For speech recording a wireless microphone system is employed. The robot constructor consists of two robot arms that act on the table. It can grasp objects, screw or plug them together, and can put them down again. The human instructor is assumed to be naive, i.e. he or she has not been trained on the domain. Therefore, speakers tend to use qualitative, vague descriptions of object properties instead of precise technical terms.

Speech recognition and shallow understanding is realized by a statistical speech recognizer (Fink 1999) that tightly interacts with the understanding component by means of a partial parsing component (Brandt-Pook *et al.* 1999). Verbal object descriptions may consist of type (bolt, cube, disc, etc.), color (red, blue, dark, etc.), size (small, long, etc.), and shape (round, angular, elongated, etc.) attributes. Out-of-domain nouns (e.g rotor) are mapped onto an unspecific *Object* label that refers to aggregated objects

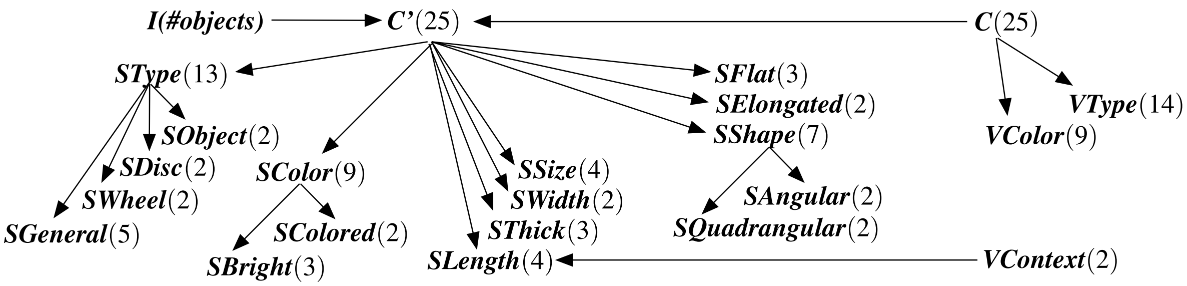


Figure 4: A Bayesian network modeling object classes. The number of states for each node is denoted in brackets.

with a higher probability. Additionally, the user can specify projective spatial relations (left-to, in-front-of, . . .), locations (in-the-middle, . . .), and structural relations (e.g. *the cube with the bolt*).

The object recognition component is based on a segmentation of homogeneous color regions. Clusters of elementary objects are parsed into aggregate structures using a semantic network (Bauckhage *et al.* 1999). The visual features used in the Bayesian network are the recognized color and type attributes of an object. Regions that do not match any object class are labeled as *unknown*.

Domain modeling

In order to capture the language use of unexperienced human instructors, a series of experiments were conducted (Socher, Sagerer, & Perona 2000). From the first three experiments frequently named color, shape, and size adjectives were extracted that are presented in Fig. 5.

gelb *(yellow)*	rund *(round)*	lang *(long)*
rot *(red)*	sechseckig *(hexagonal)*	groß *(big)*
blau *(blue)*	flach *(flat)*	klein *(small)*
weiß *(white)*	rechteckig *(rectangular)*	kurz *(short)*
grün *(green)*	dünn *(thin)*	breit *(large, wide)*
hell *(light)*	länglich *(elongated)*	hoch *(high)*
orange *(orange)*	dick *(thick)*	eckig *(angular)*
lila *(violet)*	schmal *(narrow)*	

holzfarben *(wooden)*	mittellang *(medium-long)*
rautenförmig *(diamond-shaped)*	mittelgroß *(medium-sized)*

Figure 5: Frequently named color, shape, and size adjectives.

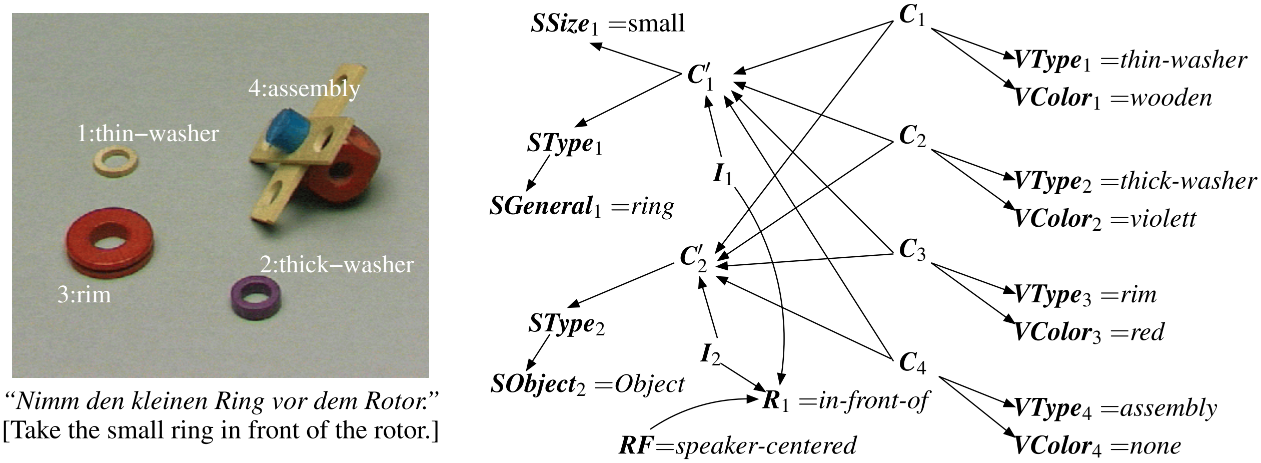
A forth experiment explored the semantics of shape and size adjectives:

- *Exp. WWW*: 426 subjects participated in a multiple choice questionnaire (274 German version, 152 English version) that was presented in the World Wide Web (WWW). Each questionnaire consisted of 20 WWW pages, one page for each elementary object type of the domain. In one version the objects were shown isolated, in another version the objects were shown together with other context objects. Below the image of the object 18 shape and size adjectives were presented that have been extracted in the previously mentioned experiments. The subject was asked to tick each adjective that is a valid description of the object type.

The evaluation of this experiment provides frequencies of use for the different adjectives with regard to the object class and the object context. A qualitative evaluation from a psycholinguistic perspective has been performed by Constanze Vorwerg (Vorwerg 2001). She extracted the following results:

1. All attributes except 'rund' *(round)* depend on context. But the context only partially determines the selection of it. In the construction kit, there exist three different types of bars with different lengths: a three-holed, a five-holed, and a seven-holed bar. The frequency of the selection of 'kurz' *(short)* and 'lang' *(long)* decreases/increases with the length of the bar as expected. This can be observed independent of the context. However, the isolated naming of a three-holed bar yields a higher frequency of 'mittellang' *(medium-long)* than that of a five-holed bar. In the context with all three types of bars present this ordering is switched. The average selection from the context version rates similar to the isolated selection.
2. The attribute selection in the corresponding dimensions, e.g. 'long' in the dimension *size*, is very specific to the object classes. Context objects have only a small influence. For example, the longest bolt is called 'long' although it has a smaller length then the shortest bar. This is not affected by the fact that there is a bar in the context or not.
3. 'dick' *(thick)* is negatively correlated with the length of an object. The bolts have all the same width, but the shortest bolt is called 'thick' with a much higher frequency.
4. 'eckig' *(angular)* is neither a super-concept of 'rechteckig' *(rectangular)* nor 'viereckig' *(quadrangular)*. It is partially used as an alternative naming.
5. 'rechteckig' *(rectangular)* is negatively correlated with 'lang' *(long)*, 'länglich' *(elongated)* is positively correlated with it.
6. Even the selection of qualitative attributes, like 'eckig' *(angular)*, depends on the context. For example, the less objects with typical angular shape are present, the more frequent 'angular' is selected.

Altogether, it reveals that the meaning of shape and size attributes is difficult to capture. It is particularly difficult to di-



"Nimm den kleinen Ring vor dem Rotor."
[Take the small ring in front of the rotor.]

Figure 6: Bayesian network for integrating speech and image interpretations

rectly extract the applicability of such attributes from image features. The solution that has been applied in this system is to use object class statistics. A first approach was proposed in previous work of Gudrun Socher (Socher, Sagerer, & Perona 2000). However she did not consider the mentioned qualitative results in the structure of the Bayesian network nor the naming of assembled objects.

The model for naming of object classes is shown in Fig. 4. *VType, VColor* represent evidential nodes containing the visual classification results of an object in the scene. *VContext* distinguishes two different contexts as discussed in item (1.) of subsection **Domain modeling**: all three different bars present or not present. We abstract from other context dependencies because this is the only one which qualitatively changes frequency orderings. The arc from C to C' depends on the value of the corresponding selection variable I. Besides the elementary object classes, the states of the variables C, C' include one state for assembled objects and an additional state 'unknown' for inconsistent object descriptions.

The evidential variables in this network are not only leaves because some nouns and adjectives are more precise than others. The meaning of *angular* and *quadrangular* were split into a super-concept modeled by the variables *SAngular* and *SQuadrangular* and *other-angular* and *other-quadrangular* states in the variable *SShape*. Object names that are not known to denote an elementary object are instantiated in the variable *SObject*. It is an abstraction of the variable *SType* and denotes an assembled object with a higher probability than an elementary object type.

The conditional probability tables (CPTs) of the object-class model are partially set by hand and partially estimated from data:

- The CPTs $P(VType|C), P(VColor|C)$ are estimated using labeled test sets consisting of 11 images with 156 objects in the first case and a pixel-based evaluation of 27 images from 9 different scenes in the second case.
- The CPTs for named object types and colors were set by hand.
- The CPTs for named shape and size adjectives have been estimated using data from *Exp. WWW*.

Spatial modeling

Besides the object class descriptions discussed in the previous subsection, spatial relations can be exploited to constrain the selection of an intended object:

> *"Take the X-object* **in front of** *the Y-object."*

The spatial model is treated as a black box function. It provides a *degree of application* for each object pair specifying if the named relation holds or not. Currently, the six different projective relations *in-front-of, behind, left-of, right-of, above, below* are distinguished. Instead of a complex 3-d spatial model, a 2-d spatial model was developed that considers each projective relation as a 2-d vector that has been projected onto the image plane (Wachsmuth 2001). The treatment of 3-d relations in two dimensions has the advantage that an error-prone 3-d reconstruction of the scene objects is not needed and complex 3-d shapes can be considered as more simple polygon areas. Additionally, a 2-d model accounts to the fact that nearly all objects are placed on the same table plane. The applicability of a relation depends on the positions of the objects relative to each other as well on the selected reference frame of the speaker. This may be either set by default, known from the context, known with a certain probability, or may remain unconstrained. All four cases can be modeled by introducing a random variable *RF* denoting different possible reference frames. The presented system distinguishes between a *speaker-centered* and a *hearer-centered* reference frame that are defined oppositely. In order to compute $P(R = \textit{in-front-of}|RF, I_1, I_2)$ from visual data, the degree of applicability of the spatial relation has to be transformed to a pseudo-probability. This is performed by calculating the degree of applicability of the inverse 2-d direction, the two orthogonal directions, and normalizing the computed degrees.

An example instantiation of the whole Bayesian network including four detected visual objects and a verbal instruction mentioning two objects related by *in-front-of* is shown

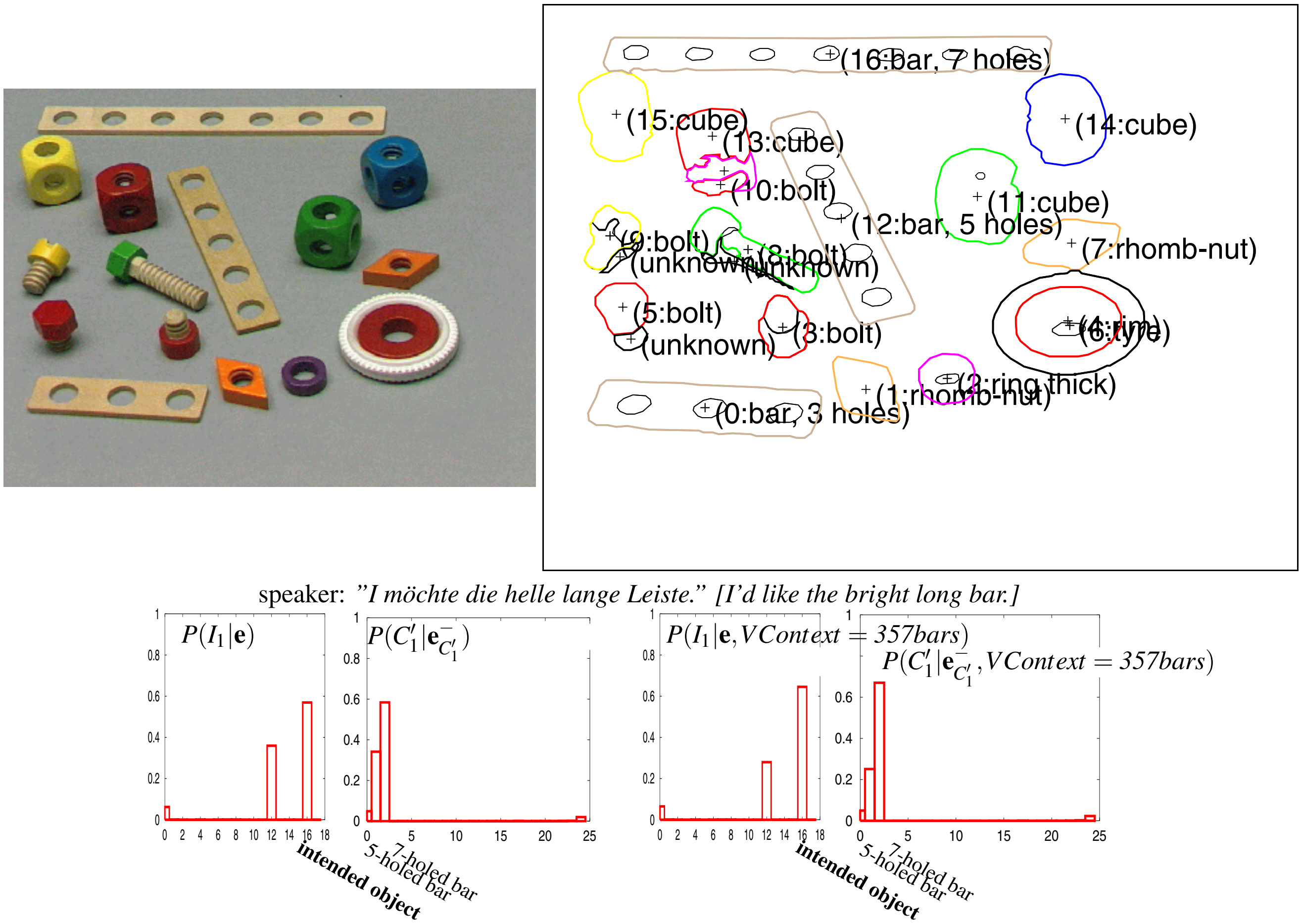


Figure 7: Performance example: The speaker intends object 16. It is correctly recognized: $\mathbf{e}^-_{C_{16}} = \{$*wooden, 7-holed-bar*$\}$. The evidences extracted from speech are $\mathbf{e}^-_{C'_1} = \{$bright, long, bar$\}$. $\mathbf{e} = \bigcup_{i=0...18} \mathbf{e}^-_{C_i} \cup \mathbf{e}^-_{C'_1}$. The first two plausibility vectors have been calculated without considering the context that all three different bar types are present in the scene. The system selects objects 12 and 16. The next two plausibility vectors have been calculated considering this context. Only the correct object 16 is selected.

in Fig. 6. Evidential variables that are not instantiated are not shown due to readability reasons. Note that the object name *ring* is an abstraction of different elementary object types and that the system has no specific semantics for the object name *rotor*. Therefore, the unspecific variable *SObject* is instantiated that indicates the denotation of an assembled object. Although the instruction is very unspecific the system is able to determine the correct intended object *2:thick-washer*.

Results

Results have been computed on a test set of 447 utterances from 10 different speakers. 10 different scenes including between 5 and 30 objects were captured by a camera and presented to them on a computer monitor. One object was marked and the speakers were asked to name it. The word error rate (WER) of speech recognition is 32%. Nevertheless the features describing an object achieve an error rate of only 15%. This is mainly an effect of the partial parser that is integrated into the speech recognition process (Wachsmuth, Fink, & Sagerer 1998; Brandt-Pook *et al.* 1999). The object recognition results include an error rate of 25% (false positive + false negative object detection + false type/color classifications). In detail, the correctly detected but misclassified objects include 18% false type classifications and 5% false color classifications.

The task of the system is to determine the marked object given the speech and object recognition results. Many speakers describe the object class instead of the individual marked object. Thus, the system is allowed to select additional objects of the same object class besides the marked one. The criterion for selecting more than one object is based on the belief vector of variable I_1. If the difference between the highest belief component and the next highest component is greater than that to the third one, a single object is selected. Otherwise additional objects are selected.

In Fig. 7 an **example of the test set** is shown that includes a relevant context switch consisting of the presence of all three different kind of bars:

1. If this context is not considered the verbal description *"the bright long bar"* denotes a *seven-holed-bar* with highest probability and a *five-holed-bar* with significant probability. As a consequence both scene objects 16+12 have significant support and are considered in the system answer, i.e. *one additional object is selected.*
2. If the context is considered in the Bayesian network the attribute *long* supports the hypothesis of a *seven-holed-bar* more strongly. Thus, the *five-holed-bar* object in the scene is supported less and is not considered in the system answer, i.e. *no additional objects are selected.*

Besides the qualitative analysis of exemplary system results, a **quantitative evaluation** was performed. From the experimental setting the following general system behavior will be expected:

- If all objects in the scene and the spoken instruction are correctly recognized, *the identification rate of the system should be high.* Nevertheless, some system answers may fail because the speaker gave a too unspecific instruction, or selected an unusual wording (e.g. some bolts were called nuts).
- If some features that were mentioned by the speaker are misrecognized or not recognized at all, *the identification rate will decrease by a similar rate as the feature error rate.*
- If some of the marked objects have been misrecognized, *the identification rate will decrease by a similar rate as the recognition error rate.*

The system results are presented in table 1. A feature error rate of 15% decreased the identification rate by only 5%. An object recognition error rate of 20% on the marked objects affected the identification rate only by a 7% decrease. The combined noisy input results in a decrease of the identification rate by 11%. Thus, the integration scheme is able to exploit redundancy coded in the combined input in order to perform correctly.

	correct input	noisy speech	noisy vision	noisy input
error rates	-	15%	20%	15%+20%
ident. rates	0.85	0.81	0.79	0.76
decrease	-	5%	7%	11%

Table 1: Identification rates

Conclusion

A Bayesian network scheme for the integration of speech and images was presented that is able to correlate spatial and type information. A universal solution of the correspondence problem was developed that can be efficiently evaluated by combining bucket elimination and conditioning techniques. The approach was applied to a construction scenario. The network structure considers qualitative results from psycholinguistic experiments. The conditional probabilities were partially estimated from experimental data. Results show that spoken instructions can robustly be grounded into perceived visual data despite noisy data, erroneous intermediate results, and vague descriptions.

References

Bauckhage, C.; Fritsch, J.; Kummert, F.; and Sagerer, G. 1999. Towards a Vision System for Supervising Assembly Processes. In *Proc. Symposium on Intelligent Robotic Systems*, 89–98.

Bauckhage, C.; Fink, G. A.; Fritsch, J.; Kummert, F.; Lömker, F.; Sagerer, G.; and Wachsmuth, S. 2001. An Integrated System for Cooperative Man-Machine Interaction. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 328–333.

Boutilier, C.; Friedman, N.; Goldszmidt, M.; and Koller, D. 1996. Context-Specific Independence in Bayesian Networks. In *Proc. of the 12th Annual Conf. on Uncertainty in AI (UAI)*.

Brandt-Pook, H.; Fink, G. A.; Wachsmuth, S.; and Sagerer, G. 1999. Integrated recognition and interpretaion of speech for a construction task domain. In Bullinger, H.-J., and Ziegler, J., eds., *Proceedings 8th International Conference on Human-Computer Interaction*, volume 1, 550–554.

Brondsted, T.; Larsen, L. B.; Manthey, M.; McKevitt, P.; Moeslund, T.; and Olesen, K. G. 1998. The Intellimedia Workbench – a Generic Environment for Multimodal Systems. In *Int. Conf. on Spoken Language Processing*, 273–276.

Dechter, R. 1998. Bucket elimination: a unifying framework for probabilistic inference. In Jordan, M. I., ed., *Learning in graphical models*. Dordecht, NL: Kluwer Academic Publisher.

Fink, G. A. 1999. Developing HMM-based recognizers with ESMERALDA. In Matoušek, V.; Mautner, P.; Ocelíková, J.; and Sojka, P., eds., *Lecture Notes in Artificial Intelligence*, volume 1692, 229–234. Berlin: Springer.

Intille, S. S. 1999. *Visual Recognition of Multi-Agent Action*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Jackendoff, R. 1987. On Beyond Zebra: The relation of linguistic and visual information. *Cognition* 26:89–114.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Francisco, California: Morgan Kaufmann.

Rimey, R. D., and Brown, C. M. 1994. Control of Selective Perception Using Bayes Nets and Decision Theory. *International Journal of Computer Vision* 12(2/3):173–207.

Socher, G.; Sagerer, G.; and Perona, P. 2000. Bayesian reasoning on qualitative descriptions from images and speech. *Image and Vision Computing* 18:155–172.

Srihari, R., and Burhans, D. 1994. Visual Semantics: Extracting Visual Information from Text Accompanying Pictures. In *Proc. of the Nat. Conf. on Artificial Intelligence (AAAI)*, 793–798.

Srihari, R. K. 1994. Computational Models for Integrating Linguistic and Visual Information: A Survey. *Artificial Intelligence Review* 8:349–369.

Takahashi, T.; Nakanishi, S.; Kuno, Y.; and Shirai, Y. 1998. Helping Computer Vision by Verbal and Nonverbal Communication. In *Int. Conf. on Pattern Recognition*, 1216–1218.

Vorwerg, C. 2001. Kategorisierung von Grössen- und Formattributen. In *Posterbeitrag auf der 43. Tagung experimentell arbeitender Psychologen*.

Wachsmuth, S.; Fink, G. A.; and Sagerer, G. 1998. Integration of parsing and incremental speech recognition. In *Proceedings of the European Signal Processing Conference (EUSIPCO-98)*, volume 1, 371–375.

Wachsmuth, S. 2001. *Multi-modal Scene Understanding Using Probabilistic Models*. Ph.D. Dissertation, Bielefeld University. http://www.UB.Uni-Bielefeld.DE/index/abisz.htm.