Multi-modal Scene Understanding Using Probabilistic Models

Sven Wachsmuth

Dipl.-Inform. Sven Wachsmuth
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld

email: swachsmu@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.).
Vorgelegt von Sven Wachsmuth am 11. Juli 2001
an der Technischen Fakultät
der Universität Bielefeld.

Gutachter:

Prof. Dr.-Ing. Gerhard Sagerer
Prof. Dr. Sven Dickinson

Prüfungsausschuß:

Prof. Dr.-Ing. Gerhard Sagerer
HD Dr. Enno Ohlebusch
Prof. Dr. Helge Ritter
Dr. Bernhard Jung

Gedruckt auf alterungsbeständigem Papier ◦◦ ISO 9706

Multi-modal Scene Understanding Using Probabilistic Models

Dissertation zur Erlangung des Grades eines Doktors der

Ingenieurwissenschaften (Dr.-Ing.)

der Technischen Fakultät der Universität Bielefeld

vorgelegt von

Sven Wachsmuth

Acknowledgments

First, I would like to especially thank my advisor, Prof. Gerhard Sagerer, for his open-minded support, continual encouragement, and fruitful discussions. His way of mediating experiences beyond technical aspects of research was very inspiring for me. I enjoyed many joint conference visits with him and the cooperative organization of the workshop "Integration of Speech and Image Understanding" which was associated with the International Conference of Computer Vision 1999.

Secondly, I very much acknowledge Prof. Sven Dickinson from the University of Toronto for being the second reviewer in my thesis committee. His comments encourage me to continue research in this area. I have to thank Enno Ohlebusch for his offhanded readiness to substitute Prof. Dickinson during my defence who had to cancel his flight because of the terrorist attack happening in New York, Sept. 2001.

The present thesis was embedded in the Collaborative Research Center 360 "Situated Artificial Communicators" that has been funded by the German Research Foundation. I joined the research project "Interaktion sprachlicher und visueller Informationsverarbeitung" (*Interaction of speech and image processing*) for three exciting and inspiring years.

The work would not have been possible without the support of my colleagues of the "Applied Computer Science" group at Bielefeld University. The intense social and scientific interaction within the group has been very fruitful for my own work. Franz Kummert, Gernot Fink and Stefan Posch provided many useful comments, hints, and ideas during the last four years. I have to especially thank Gernot Fink, Hans Brandt-Pook, Franz Kummert, Christian Bauckhage, and Jannik Fritsch who provided the program modules that realize the basic speech understanding and vision capabilities of the system. My office partner Frank Lömker could answer nearly any technical question concerning implementation issues. Finally, I was able to start from previous work and to use data collected by Gudrun Socher, Constanze Vorwerg, and Thomas Fuhr.

Furthermore, I very much acknowledge Gernot Fink and Thore Speck. Their comments on my thesis significantly improved the quality of the presentation both in content and language.

Last but not least, I like to thank my family Beatrix and Joris for their patient and support. The birth of Joris during the preparation of my thesis gave me the necessary motivation and energy to finish it.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

How do we explain a picture to another person? We talk about the picture, describe the colors, shapes, and objects in it, mention how different objects are related to each other. How do we explain a verbal statement? We show a picture which visualizes the content of the utterance, the objects mentioned in it and how they are related. In everyday communication people use various ways in parallel in order to transmit their intention. They point on something, put on a special face, gesticulate, or refer to the common environment of the communication partners. They use different ***modalities*** in order to communicate. It seems to be just natural to use the same way of interaction in ***human-computer-interfaces***. The consequence is a paradigm shift from passive interfaces, such as mouse clicks or text typing, to an active communication partner that interprets the auditive and visual environment, draws inferences using background knowledge, and requests missing information. Subsequently, such an active human-computer-interface will be called ***artificial communicator***.

However, the automatic interpretation of signals of a separate input modality, such as speech understanding, gesture recognition, or visual object recognition are only one part of the total. In order to build systems which communicate with people in a natural way, the integration of modalities is an essential task that is not trivial. Each modality has its own vocabulary and expressiveness. Pointing defines a region or direction of interest, a special face may represent an emotional feeling, speech understanding provides qualitative facts about the world, and vision perceives and interprets analogous shapes in the world. I think it is not questionable that different formalisms are needed for processing different modalities, and, indeed, this is the fact in the current state of the art (see Sec. 2.2,2.3). The question is, and this thesis will be an experimental study in this topic, what is the most promising formalism to integrate the results of the specialized processing components of such a multi-modal system or artificial communicator? How should the individual components of the system be connected, and how should the processing be organized? This thesis will give an innovative answer to these questions and present a realization in a particular domain.

The most complex modalities for human-human communication are the use of language and the visual perception of the environment. The understanding of speech and visual impressions are abilities that people use extraordinarily well. But they do even better if they have the chance to integrate visual and auditive information. Many psychological studies have supported this theory, e.g. [BJ72, LP74]. The simulation of the visual and auditive capabilities of human beings on a computer has achieved impressive results. But the generality and complexity of successfully interpreted signals is far from that accomplished by human beings. Therefore, such research is still an extreme and inspiring challenge. Due to the overwhelming complexity of these understanding tasks several disciplines have been established in the science community. ***Computer Vision*** (CV) is concerned with the interpretation of images and image sequences. ***Speech Recognition*** (SR) is the task of mapping from a digitally encoded acoustic signal to a sequence of words. ***Natural Language*** (NL) processing takes a sequence of words and extracts the intention of the sentence. They all have achieved impressive results by developing their own techniques (cf. e.g. [BB82, RJ93, All95]).

However, the need for integration of different modalities, especially for NL and CV, has been proposed by various authors [Win73, Wal81, Jac87, MP96]. In a very influencing paper of 1973, Terry Winograd states that

> "*Most attempts to model language understanding on the computer have followed this strategy of dealing with a single component of language.*" [Win73, p. 152].

Although Winograd did not have CV in mind, he "describes an attempt to explore the interconnections between the different types of knowledge required for language understanding." [Win73, p. 153]. He describes a system, SHRDLU, that can answer questions and follow instructions in a simple world of blocks. It uses an internal analogous world representation that provides the context of the discourse understanding. Therefore, SHRDLU is one of the first systems integrating natural language understanding and, in this case, an internal representation of a virtual environment. Twenty years after the publication of Winograd Hans-Hellmut Nagel asks:

> *"Why did it take so long to extend SHRDLU or analogous germs into approaches which evaluate image sequences in order to derive an estimate of the actual configuration of objects in a depicted real world scene and to represent the recorded activities of (certain) entities as agents by natural language concepts?"* [Nag94, p. 98].

There seems to be inherent difficulties in switching from an internal virtual environment to an external realistic environment that is observed by a camera. Nagel mentions the complexity of real world scenes, the sheer mass of data to be evaluated (even in the simplest case of monocular gray value video sequences), and a lack of experience in the development and exploitation of programming tools to cope with the aforementioned problems as the main reasons. And I will add that one of the most serious problems in relating results of a vision system and a speech understanding system is the treatment of uncertainties. In computational systems these uncertainties have to be represented in numbers leading to the question: *How to get those numbers?* Even today, despite several very successful research projects and innovating approaches, the question how to relate natural language and image concepts is generally unsolved. Should we use neural nets as Steven Harnad suggests [Har90], a fuzzy temporal logic like that of Hans-Hellmut Nagel

[Nag99], or may a probabilistic framework like Bayesian networks [Jen96], which have become very popular over the last ten years, provide the best concepts? I think it will be an open race even for many years in the future. The only way to obtaining new advances in this field is to build such systems and to reduce the lack of experience mentioned by Nagel.

This thesis represents a new step towards a reduction of this lack. It introduces the usage of Bayesian networks in the field of integrating speech and image understanding. In an assembly scenario consisting of a robot constructor and a human instructor, an essential part of a human-computer interface is realized which is able to establish referential links between the visually observed scene and the spoken instructions of a user. This is a work in progress that has been started by Gudrun Socher [Soc97] and was supported by the work of many other colleagues in the *Applied Computer Science* group at Bielefeld University. Socher presented a first step towards employing Bayesian networks in a speech and image integrating task. In the course of my own work, the paradigm of the integration component changed from a hybrid approach towards a unified Bayesian network approach. This will be discussed in more detail in chapter 4.

In the following chapters the contributions of this thesis are presented in detail. **Chapter 2** gives an introduction to the topic of integrating verbal and visual information. After a short review of basic principles in computer vision and automatic speech understanding, different directions of research are discussed leading to a positioning of this thesis. **Chapter 3** developes the theoretic background of the proposed solution. Bayesian networks are introduced as a mathematical framework for reasoning with uncertainties. **Chapter 4** applies the theoretical model to a particular domain. A human-computer interface is realized for a construction scenario. **Chapter 5** examines different possibilities of drawing inferences in the proposed uncertainty model giving several performance examples of the implemented system. Further learning issues are discussed as an outlook. **Chapter 6** quantitatively evaluates the realized system showing its roboustness with regard to erroneous input data. Finally, **chapter 7** summarizes the contributions of this thesis and gives an outlook to future work.

Notational remarks

Names, spoken text, and emphasized words or passages are written in *italic*. When a new term introduced in the following text this is written in ***bold-face italic***.

Chapter 2

Problem Statement

Abstract. *In order to understand the contribution of this thesis, the positioning and the limitations of the solved problems must be known. Therefore, an overview of the research directions concerning the integration of speech/NL and image processing is given and some basic principles of automatic speech understanding and computer vision as separate modalities are presented. Finally, the scope of the thesis is described in more detail and related work from literature is discussed.*

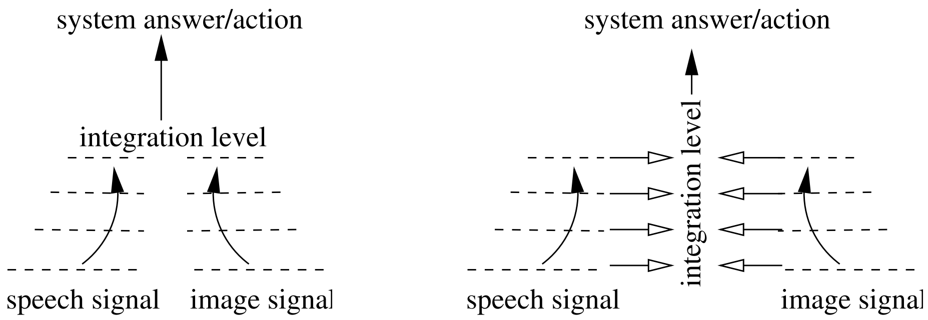
2.1 Robust Processing in Human-Computer Interaction

The development of intuitive and natural human-computer interfaces is an important research topic that influences and is influenced by the integration of computer systems in our daily life. The perception of what we say and what we see are two natural abilities that constitute an intuitive communication. However, the natural character of a communication does not exclusively rely on the kind of modality that is used. It also depends on the syntax and processing of the transmitted content. An instruction like "`Computer - take - object - eleven`" is very unnatural in the choice of vocabulary, in terms of speaking isolated words, and using unique numbers for object identification. Its syntactic structure is very simple and well defined. A natural instruction would be for example "*Mhm now I'd like you to take this big/ er long part that is in front of – the Phillips screwdriver*". The words may be spoken with varying speed, words are shortened ("*I'd*") or aborted ("*big/*"), and hesitations ("*Mhm*", "*er*") and corrections violate the syntactic constraints. The identification of objects suffer from vague descriptions ("*long ... in front of*"). Words like "*Phillips*" might not be modeled. A big problem is the selectivity of any choice of vocabulary. Furnas et al. showed that for complex systems "many, many alternative access words are needed for users to get what they want" [FLGD87, p. 971]. First, they informally describe an experiment during which people were asked to name a command for a specific information retrieval, namely interesting activities in a major metropolitan area. Less than a dozen pairs of more than a thousand selected the same name. This is confirmed by six experiments in five different domains: text editing (48 typists), decoder command naming (100 system designers), common objects (337 students), category selection of ad items (30 homemakers), and keywords for recipes (8 cooks, 16 homemakers). The probability of two people applying the same term to an object is reported to be between 0.07 and 0.18 [FLGD87]. They conclude that an interface design has to "begin with user's words, and find system interpretations. We need to know, for every word that users will try in a given task environment, the relative frequencies with which they would be satisfied by various objects." [FLGD87, p. 968].

In the same manner, for a vision system isolated symbols or bar codes, colored line drawings or geometric solid blocks can be recognized more easily than pliers and numerous other tools in a complex workshop environment. A two-class problem can be solved more accurately than the distinction of a hundred different object classes or the localization and recognition of general shapes. Experimental settings with controlled lighting conditions can be handled more stablely than outdoor scenes with different weather conditions. An unconstrained system that is able to recognize and understand general speech, to detect, localize and classify general objects, and understand its natural visual environment is currently not feasible. The task of a researcher is to find a compromise between complexity and tractability. Therefore, the term ***robust processing*** is always limited to a constrained test domain. Anything outside this test domain will be treated as noise. In this sense, one aspect of robustness is the performance of the system under noisy input.

However, besides this external source of noise, the system has to face an inherent error source resulting from the ability of the system to abstract from measured data. The processing of such audio or image signals is typically divided into different processing levels in order to simplify the processing task (see the following sections 2.2, 2.3 for a further discussion). For example, the lowest level represents a digitized image. The next one extracts regions of homogeneous texture or color. On the following level the shape, texture, and color of those regions is characterized by a set of classification numbers (*eccentricity, spatial frequency, color values*). One step further, combinations of shape descriptions are assigned to object labels (*adjacent wooden and metallic elongated regions with orthogonal main axes may form a hammer*). The highest level may qualitatively describe the relations of the objects found in the observed scene (*the hammer is in the toolbox*). In traditional horizontal architectures the system answer or action is solely based on the highest level of interpretation (Fig. 2.1(a)). However, on each processing level decisions reduce the information basis of the subsequent level. Regions abstract from specific pixel values. Shape descriptions abstract from a precise boundary representation. Object labels abstract from the appearance of the image object. Qualitative descriptions abstract from the numeric coordinates of the position and orientation of an object. Consequently, any decision on a lower level deeply influences subsequent decisions on higher levels. Errors are propagated. The information that *'the hammer is in the toolbox'* relies heavily on the decision if some pixel values have the same color.

The ***robustness*** of the system describes how the system answer or action is affected by propagated errors. Decisions on lower processing levels typically consider local characteristics of the input signals that may indicate a wrong hypothesis. Therefore, the key issue for a robust system is to passively or actively control the subsequent reduction of the input data to a qualitative symbolic description. Temporary inconsistencies may be resolved by exploiting the ***redundancy*** in the verbal coding of the speaker's intention, in



(a) High-level integration

(b) Multi-level integration

Figure 2.1: Traditional high-level integration schemes (a) lose much information because of low level errors that are propagated to higher levels of abstraction. Integration is performed without considering the cause of errors on higher levels. Multi-level integration schemes (b) are able to model these error propagations which consequently leads to a more robust system behavior.

the visual coding of a referenced object, and the redundancy in the combined auditive and visual information. As a consequence, high-level integration schemes must be substituted by multi-level integration schemes (Fig. 2.1(b)) that open up the possibility of a tighter interaction between the interpretation processes of different modalities.

2.2 Basic Principles of Computer Vision

Since the early days of artificial intelligence in the nineteen fifties the idea of building a machine that can perceive its visual environment just the way like humans can see using their eyes has been a great challenge for computer scientists. Various approaches have been proposed and many books have been published about this topic (cf. e.g. [BB82, Fau93]). Since a general discussion of this topic will be far beyond the scope of this thesis, this section concentrates on some basic principles which can be found in almost every computer vision system. This section is mainly based on an introductory chapter by Sven J. Dickinson [Dic99], and a more cognitive overview by Steven Pinker [Pin84]. Dana Ballard and Christopher Brown define ***Computer Vision*** (CV) as follows [BB82, p. xiii]:

> ***Def.:*** *Computer Vision is the construction of explicit, meaningful descriptions of physical objects from images.*

Physical objects may be any kind of entity which is relevant for an application. For example, if we wanted to find ships in aerial images showing a harbor, the physical objects would be the ships lying in the harbor and the description would be their type and position. In a medical application checking the functionality of the heart valves these would be the physical objects and the description would include the regularity of their movement. In

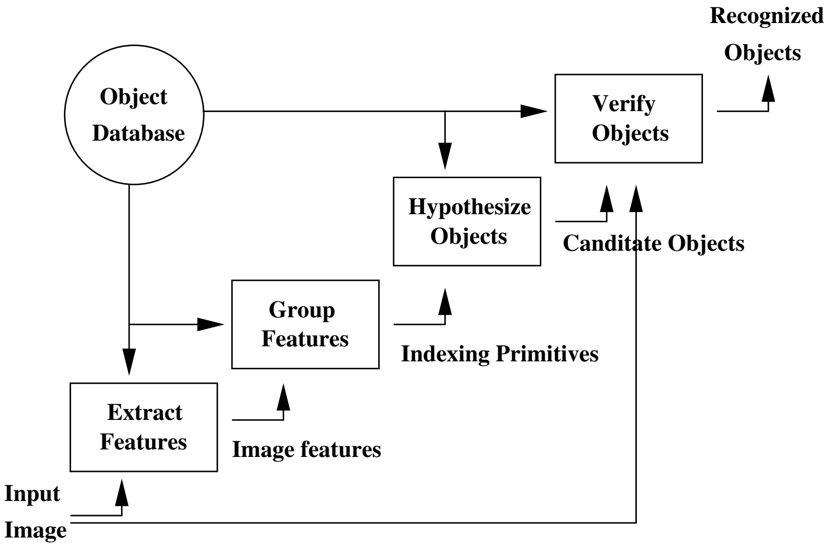


Figure 2.2: Components of an object recognition system (cf. [Dic99]).

the first case, a single still image – maybe infra-red – will be interpreted, in the second case a sequence of ultrasonic images will be the input for the computer vision system.

This section will concentrate on some aspects concerning the ***object recognition*** task, i.e. assigning a label to an image object. It is one of the primary functions of the human visual system. Recognition allows us to understand the content of images and to ground the image object in our own experience [Dic99]. In a computer vision system the label assigned by object recognition links the pixel image of a physical object with the knowledge base of the system. In Fig. 2.2 typical components of an object recognition system are shown. The input of the system is a ***digital image*** and an ***object database*** containing a number of ***object models*** that code the knowledge of the system about the objects to be recognized. In the first step an appropriate set of ***features*** are extracted from the image such as edges (brightness discontinuities) or regions (homogeneous image patches). These features may be *grouped* in order to *hypothesize objects* in the image. An inherent task of the grouping process is the separation of an image object from its background, because the grouped collection of features is used as a search key or ***indexing primitive*** in the object database and, therefore, should only contain information caused by one object. The generation of object hypotheses is a matching procedure of grouped features and object models. If a single object model has a valid match the recognition procedure is complete. Otherwise a last step has to verify each object hypothesis (matched object model) in the input image in order to decide for the most probable one.

The first three steps are described as bottom-up (data-driven) processes, whereas the verification step is a top-down (model-driven) process. However, this does not imply that the grouping process may not influence the feature extraction process or that the grouping is independent of the selection of candidate objects. These very complex interactions may certainly increase the reliability and performance of the recognition system. The ***bottom-***

up and ***top-down*** terms shall indicate the initial starting point of the analysis process.

In order to make the different processing steps of this abstract object recognition scheme more concrete, the following passages will describe how different recognition paradigms fit into this framework.

Template matching is an aspect or view-based scheme. The object database contains replicas of digital image projections of the physical object stored in an object model. The feature extraction is just the selection of an appropriate window of the image or the image itself. This is used without any grouping directly as an index to the database. The matching is mostly realized by computing a correlation measure and the object model with the highest correlation is selected.

In ***feature models*** an image object is transformed to a finite set of classification numbers, one for each feature (feature extraction and grouping). A feature may be thought of as a "mini-template" that characterizes the shape of the image object. If an object is to be recognized from different positions, angles, distances, or in different lighting conditions, the features should be invariant across the corresponding transformations such as translation, rotation, scaling, or variation of illumination. The classic approach to realize the indexing into a finite set of object models given the set of features is ***pattern classification*** [Sch96]. Each object model corresponds to one object class κ. The object image is interpreted as a pattern $\vec{f}(\vec{x})$ that describes a function of measurements taken at positions $\vec{x}$. For each pattern a feature vector $\vec{c}$ is given, that is the set of classification numbers. A discriminating function $d(\vec{c})$ maps each feature vector to a class κ, that is the index of the object model. The discriminating function may be defined by distance measures (e.g. next neighbor classification), minimization of an error measurement (e.g. back propagation networks, polynomial classifiers), or probability distributions (e.g. Bayesian classifiers) (cf. e.g. [Sch96]).

The idea of ***structural descriptions*** is a divide and conquer strategy. In order to recognize a complex object it is *divided* into meaningful subparts that are recognized more easily. If these subparts have been detected in an image they may be combined by testing specific relations stored in the object model (*conquer* step). An object model may consist of several decomposition levels resulting in a complex hierarchical description. Now, the feature extraction process has to detect primitive subparts in the image. The grouping process has to select a subset of primitives which shall be checked against the object database. Often, the object models have graph-based representations with primitives as nodes and relations between these subparts as edges. Such relations may also be established in the grouping process, so that the indexing into the object database can be realized as a comparison of two graph-based descriptions (see e.g. [GLP84],[Vos91]).

In all three cases (template matching, feature models, structural descriptions), nothing has been said about the verification step in the object recognition scheme. The reason is that any additional criterion that may be used for a verification can also be used in the indexing step because it is defined in the same formalism. A useful verification criterion must be based on another source of information or another complementary extraction method. Therefore, the object verification step is a possibility to combine different recognition paradigms, e.g. to verify an hypothesis based on a feature model, which may be computed very fast, by a structural analysis, which may be more time consuming.

The recognition of 2-d and especially 3-d objects suffers from diverse difficulties that cause recognition errors to be inherent to any object recognition strategy. Perspective displacements, dilations or contractions, rotations, parts that disappear or emerge, images that become blurred and lose finer details, or lost parts that are occluded are only some aspects which have to be considered when defining object models and indexing strategies. The correct grouping of features suffers from the problem that edges or regions can neither be distinguished from edges and surface details of surrounding objects, nor from the scratches, surface markings, shadows, and reflections of the object itself.

However, the recognition strategies have different benefits and drawbacks. Template matching works well for isolated objects and has difficulties with occlusions or changes in distance, location, and orientation. Feature models can easily take account of invariant properties of image objects, but typically do not consider spatial dependencies of model features which causes some serious problems (see [MP72]). Another restriction is that it is almost impossible to define natural shapes in terms of a fixed dimensional vector of classification numbers. The advantage of structural descriptions is the explicit representation of meaningful decompositions and relations. They can be easily used to reason about the structure of an object and to connect the object structure with background knowledge. Another possibility is the definition of generic object models which describe an unlimited set of objects by means of combination rules of primitives. The most serious disadvantages are that we need another paradigm in order to recognize the primitives of our model and that such models are very difficult to learn automatically.

A computer vision system which is used as a perceptual front-end of a human-computer interface must enable the complete communication system to reason about the visual environment in a robust manner. Especially, it has to define interaction points with other modalities like speech understanding. There are two steps in the object recognition scheme that are promising for interaction purposes:

- **The indexing step:** Visual features can be combined with verbally mentioned features in order to obtain more precise queries to the object database. Therefore, the vision system must be able to provide or represent unclassified object hypotheses that are described by some set of visual features or another kind of visual characterization. These indexing primitives will be called ***unknown objects***.

- **The verification step:** Verbal information can be used in order to detect visual *recognition errors*, to refine object categories, or to weight *competing interpretations*, i.e. alternative object hypotheses. Without previously calculated visual information, the object model that is hypothesized by the verbal information can define a *top-down starting point* for visual analysis.

This thesis will explore aspects of both possibilities, which may be considered to happen on different layers of abstraction. It will be shown that both kinds of interaction can be realized in the same interaction framework.

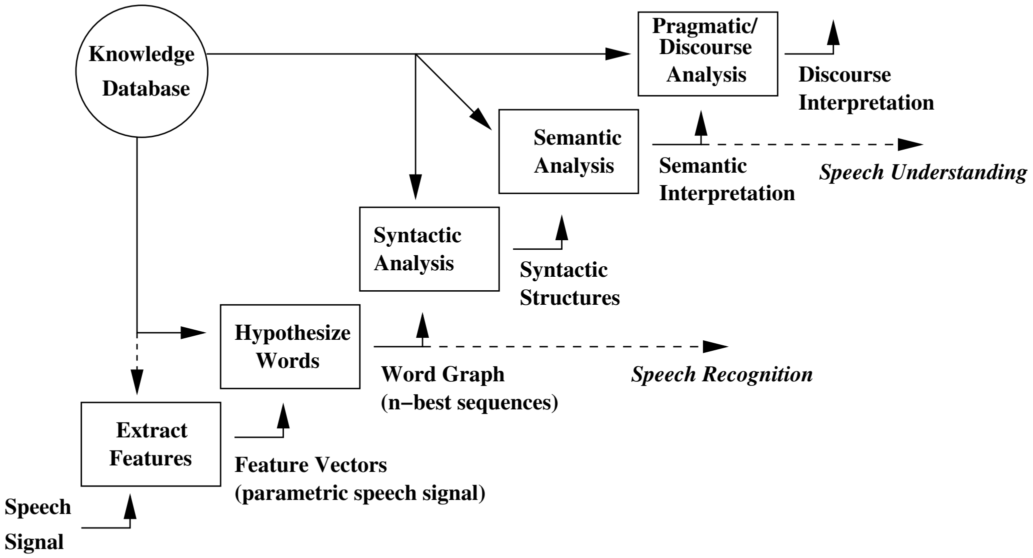


Figure 2.3: Components of a speech understanding system in a discourse context.

2.3 Basic Principles of Automatic Speech Understanding

Language is the most effective medium in human communication. However, that we effortlessly understand and use language in our daily life is in contrast to the complexity of the information processing task that has to be realized in a computer system which has only marginal language understanding capabilities.

In computer science, ***formal languages*** – like ‘C’ or ‘Lisp’ that are invented, rigidly defined, and easily processed by computers – are divided from ***natural languages*** – like Chinese, English, or German that humans use to talk to each other. In the following some basic principles and problems concerning the understanding of spoken natural language will be discussed. Traditionally, this research area is divided into ***speech recognition*** and ***natural language*** (NL) processing or understanding. Therefore, this section is also devided into two parts. The speech recognition part is merily based on the books of Kai-Fu Lee [Lee89] and Douglas O'Shaughnessy [O'S00] whereas the understanding part is based on the books of James Allen [All95] and Ray Jackendoff [Jac89].

Both reseach fields have a long history in artificial intelligence. Russel and Norvig give the following definition for speech recognition [RN95, p. 757]:

> ***Def.:*** *Speech recognition is the task of mapping from a digitally encoded acoustic signal to a string of words.*

In linguistic terms, speech recognition is a transformation from the ***acoustic level*** to the orthographic or ***textual level***. An important subtask is the segmentation of the speech stream into single utterances, words and, sometimes, syllables. These discrete linear elements are not directly apparent on the acoustic level. For example, consider the boundary between words in a pair like “Dick stops” versus “Dick’s tops”. The distinction between their acoustic representations is not a space of silence before the ’s’ or after the ’s’. The primary difference appears in the acoustic realization of the ’t’ that follows the ’s’ (see [Jac89, p. 58]). The phonetic structure and the acoustic realization of words is mostly modeled by Hidden Markov Models (HMMs) that combine the decompositional aspects with the framework of a statistical classifier. The main challenges of speech recognition are ***speaker independence***, ***continuous speech***, ***large vocabularies***, ***natural tasks*** [Lee89, pp. 2], and ***robustness*** [MRB00].

The selection of appropriate features for the recognizer is mainly based on experience, and most parametric representations of a speech signal are highly speaker-dependent. Consequently, a set of reference patterns suitable for one speaker may perform poorly for another. In order to recognize speech from any new speaker, speech parameters have to be defined relatively invariantly between speakers, or multiple representations have to be used. Another possibility is speaker adaptation which starts with an existing set of parameters that are slightly modified during the first sentences of a new speaker.

Continuous speech is significantly more difficult than isolated word recognition. First, word boundaries are unclear and, secondly, co-articulatory effects are much stronger. Thus, the pronuncation of the preceeding word ending and that of the subsequent word depend on each other. Thirdly, content words (nouns, verbs, adjectives, etc.) are often emphasized while function words (articles, prepositions, pronouns, short verbs, etc.) are poorly articulated.

In 1989, ***large vocabulary*** speech recognition started with vocabularies of about 1000 words [Lee89]. Today, this term typically means a vocabulary above $10,000$ words, e.g. Hermann Ney and Stefan Ortmanns report speech recognition results with a vocabulary of 64,000 words [NO99]. One fundamental problem of large vocabulary recognition is that words cannot be modeled individually, because of a lack of sufficient training material. Therefore, appropriate subword units have to be identified and used. The more the vocabulary is increased, the more difficult is the classification problem because the search space of possible word sequences explodes. Thus, special search strategies must be applied in order to retain control [NO99]. The word context becomes even more relevant and must be exploited in the processing as early as possible to restrict the search space and prune unpromising search paths. Typically, ***language models*** in form of grammars or probabilistic *n*-grams are used to constrain possible word sequences. However, there is a trade-off between the size of the vocabulary and the restriction on word sequences. Large vocabulary speech recognition is mostly applied to controlled read speech whereas acceptable recognition rates for uncontrained spontaneous speech are only obtained for smaller vocabularies.

The term ***natural task*** refers to these constraints on possible word sequences that may be uttered by a speaker. As mentioned previously, they are typically represented by ***language models*** in form of grammars or probabilistic *n*-grams. Besides the restriction of the search space, language models are a powerful technique to disambiguate homophones, such as ‘buy’ and ‘bye’, or words with a similar pronuncation, such as ‘dog’ and ‘dock’. One aspect of natural tasks is covered by the ***perplexity***, an information-theoretic measurement of the average uncertainty at each decision point, i.e. the word possibly uttered next (cf. e.g. [Lee89, p. 8]). Another aspect is that of novel, erroneous, or incomplete use of language. These phenomena – typically refered to as ***spontaneous speech*** in contrast to ***read speech*** – violate any strict definition of a sentence grammar and, therefore, must be handled by exception rules or soft grammar definitions.

Humans even recognize speech in the presence of noise. They are able to separate the speech signal from environmental background noise (e.g. in cars), human noises (e.g. breathing or smacking, other speaking people), or echo effects. The transferred ability of computational speech recognition systems is called ***robustness***. It describes how the performance of a system changes from laboratory to realistic environmental conditions [MRB00].

In parallel to *computer vision*, the term ***natural language understanding*** may be defined by the following task definition:

> ***Def.:*** *Natural language understanding is the construction of an explicit, meaningful representation of a string of words.*

Therefore, the interface between speech recognition and natural language understanding is a string of words. The problem with this widely used architecture is that knowledge sources and processing are both completely separated into recognition and understanding parts. These constraints may either be softened by exchanging the n-best solutions of the recognition process – that can be represented compactly by a word graph – (see e.g. [AN95, WAWB$^+$94]) or by incorporating knowledge used in the understanding part into the recognition process (see e.g. [GZ92, HW94, WFS98, BPFWS99]).

Traditionally, three levels of processing are distinguished in natural language understanding. The ***syntactic level*** describes the structure and word order of a sentence. First, words are classified into parts of speech or ***lexical categories*** such as noun, verb, adjective, adverb, preposition, and conjunction. Secondly, words are combined into phrases, which are themselves classified into ***phrasal categories*** such as sentence, verb phrase, noun phrase, and prepositional phrase [Jac89, p. 68].

The distinct treatment of the ***semantic level*** and the ***pragmatic level*** is a matter of discussion in the language processing community. "*Semantics* is the study of aspects of meaning that are due purely to linguistic form" [Jac89, p. 121]. It analyzes the meaning of a sentence independent of the context. "*Pragmatics* is the study of aspects of meaning that arise from the interaction of language with one's nonlinguistic perceptions, with one's knowledge of the social circumstances in which the sentence is uttered, and with one's general knowledge of the world" [Jac89, p. 121]. Such a context may also be given by the history of a discourse in a dialog. Both levels may be subsumed under a ***conceptual level*** which is based on a unique knowledge representation.

The automatic understanding of natural language is far from being straightforward. A central problem are various cases of ***ambiguity*** – i.e. *competing interpretations* – that appear on each level of analysis. Lexical ambiguity emerges if a word has more than one meaning ("hot": warm, spicy, electrified, radioactive, etc.) or category ("back": go back, back door, the back of the room, back up your files, etc.). Syntactic or structural ambiguity is often referenced for propositional phrase (PP) modifiers ("He sees the man with the telescope.") that may be attached to either the verb or the noun phrase. Semantic ambiguities arise as a consequence of lexical and syntactical ambiguities or different meanings of word combinations ("coast road": road follows or leads to the coast). Referential ambiguity is a pervasive form of semantic ambiguity. Typical forms are anaphoric expressions such as “it” that may be a representative for any entity. Pragmatic ambiguity results from a disagreement between the speaker and the hearer on what the current situation is (“We will meet next Friday”). Vague meanings are another kind of semantic ambiguity that may have consequences on referential ambiguities (“I will take the big one.”).

There are several evidences that can be used in order to disambiguate competing interpretations. Lexical evidences may be a preference of one meaning of a word. The preference to attach a modifier to the most recent constituent is a syntactical evidence. The semantic interpretation of other parts of the sentence introduces another kind of preference that can be formalized as a conditional probability (“ball, diamond, bat, base” in the baseball senses, “I ate spaghetti with a fork” versus “I ate spaghetti with a friend”). A pragmatic evidence can be given by information extracted from the discourse history or the visual context (we have been told that “he” is using a telescope, or we see that “the man” carries a telescope).

The usage of speech as a fast, robust, and natural input modality in human-computer interaction has consequences on the representation and processing issues. Spontaneous speech introduces many phenomena such as novel words[1], aborted words, incomplete sentences, and hesitations that force recognition errors and increase the complexity of the understanding task. The sentence structure is often corrupted so that *partial interpretations* of utterances are needed. Visual information may improve the performance of a speech understanding system on nearly every level of processing. On the orthographical or word level, visual information can be used to constrain word sequences expected by the speech recognizer. As reported in [NSF$^+$95] the perplexity of a test-set could indeed be reduced by applying such a strategy. But a significant reduction of the error rate could not be achieved. *Syntactical ambiguities* can be resolved by looking into the scene. Therefore, a scoring scheme based on visual information can be integrated into a parsing strategy [KWK99]. On a semantic level, *partial information* about referenced objects or actions can be completed by visual information. An important observation in the case of speech input, in contrast to textual input, is that a fused object description might not be consistent due to *recognition failures*. Additionally, establishing the semantic correspondence between verbal descriptions and object classes might be affected by newly introduced ***unknown words***. This thesis will refer to the term *unknown word* as a word with unknown semantics. Thus, it is included in the recognition lexicon and the possible syntactic categories will be constrained to specific open vocabulary word classes. These restrictions simplify the task of processing unknown words in order to make this problem tractable for the scope of this thesis.

[1] Here the term *novel* is used with regard to a modeling lack in the speech recognition or understanding part. In the following, this thesis will refer to the term *unknown words* as a word with unknown semantics, i.e. it is included in the recognition lexicon and the syntactic category noun is assumed.

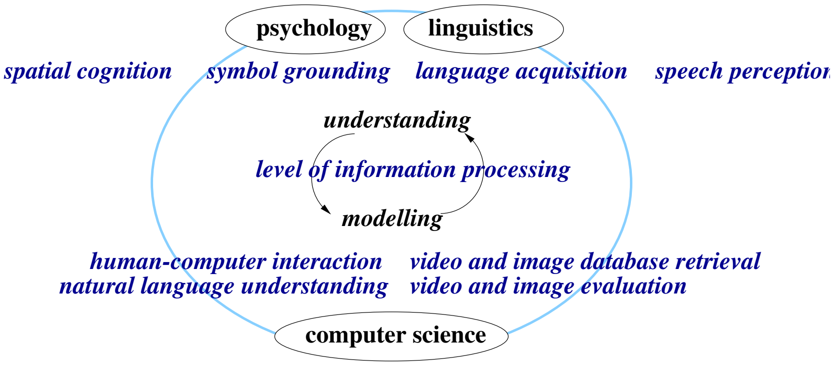

psychology
linguistics
spatial cognition
symbol grounding
language acquisition
speech perceptio
understanding
level of information processing
modelling
human-computer interaction
natural language understanding
video and image database retrieval
video and image evaluation
computer science

Figure 2.4: Research directions in the field of integrating speech/NL and image processing.

2.4 Integration of Speech and Image Processing – An Overview

The topic of integration of speech and image processing is an interdisciplinary study. The way how psychology, linguistics, and computer science interact is very well characterized by Jeffrey Mark Siskind: "I believe that the link between language on the one hand, and perception and action on the other hand, is the cornerstone of higher cognition. Understanding and modeling how we talk and reason about what we see and do is the key for us to understand our own minds and ultimately create artificial ones." [Sis98, p. 2]. The dualism of *understanding* and *modeling* is the main aspect that will be reflected in the following considerations.

2.4.1 Psychological experiments and the level of information processing

Psychological experiments that explore this link aim at answering questions about human cognition. However, experimental studies in this area have to be designed and interpreted very carefully. As Stephen M. Kosslyn states: "The scientific method rests on being able to distinguish among alternative hypotheses to everyone's satisfaction, which requires that the subject be publicly observable. [...] Mental events are notoriously difficult to put on public display." [Kos94, p.1-2]. If mental states cannot be observed and interpreted directly, the level of description has to be changed. Therefore, most psychological studies describe cognitive systems at a more abstract functional level, i.e. the level of information processing as described by Newell, Shaw, and Simon: "At this level of theorizing, an explanation of an observed behavior of the organism is provided by a program of primitive information processes that generates this behavior" [NSS58, p. 151]. The theory is not at all concerned with physical structures in the human brain. Instead, human behavior is reconstructed by a number of memories, a number of primitive information processes operating on them, and a set of rules that combine these processes into complete programs. Using this perspective, many phenomena have been described which imply a tight interaction of visual and verbal processing. For example, Paivio has discovered that one's ability to learn a set of words can be predicted well by how easily one could visualize their referents [Pai71]. John D. Bransford and Marcia K. Johnson have shown that subjects would be able to comprehend a passage of text quite easily if they received the appropriate prerequisite knowledge from a picture providing information about the context. Subjects who did not have the access to the appropriate knowledge would find the passage difficult to understand [BJ72]. In another experiment described by Elizabeth F. Loftus and John C. Palmer subjects watched films of automobile accidents and then answered questions about events occurring in the films. They show that the phrasing of the question used to elicit the speed judgment (use of different verbs: smashed, collided, bumped, hit, contacted) influences the estimate. They draw the conclusion that questions asked subsequently to an event can cause a reconstruction of that event in one's memory which influences the answers of further questions [LP74]. The ability to reconstruct an event in one's memory is further referenced by Shepard and Cooper who showed that people can mentally rotate objects in images, and that this rotation operation is incremental. Therefore, they conclude that analogous representations are used in mental reasoning [CS73, SC82]. A much deeper discussion about the representation and reconstruction of visual sensations in the brain can be found in the book "Image and Brain" of Stephen M. Kosslyn [Kos94]. He especially discusses the need of analogous representation against a pure propositional approach which is also known as the ***imagery debate***.

2.4.2 Linguistics and the symbol grounding problem

The linguistic part is very tightly coupled with psychology and computer science. The main topics which are relevant in this context are language acquisition and speech perception. The first one investigates how children are able to learn the use of language whereas the second one develops theories about the mental process of understanding language. Both areas have been influenced by methods developed in ***artificial intelligence***, which tries to build intelligent systems using computational methods. It is intensively discussed whether computer systems, in principle, are able to simulate human cognitive processes such as language understanding. Does a computer program really understand a text, does it really learn the meaning of a new word, or can it only manipulate some symbols? These even psychologically relevant questions are raised in the ***symbol grounding problem*** as described by Stevan Harnad [Har90] and Searle's Chinese room [Sea80]. The key point of both argumentations is that a purely symbolic system cannot learn and represent the meaning of its symbols. Therefore, Marconi states, that "natural-language understanding systems [which are typically realized as symbolic systems] are only metaphorically such, for they do not *really* understand natural language." [Mar96, p. 120]. In order to understand a story, Marconi distinguishes ***inferencial competence*** and ***referential competence*** of a system. The former is the ability to draw conclusions from facts that we know from the story or that had been given before, e.g. from "there are four elephants in the living-room" the system may infer that 'there are an even number of animals in the house'. The latter refers to the ability to check the truth condition of such sentences in the "real world". A symbolic system cannot verify the sentence unless the "real world" is given to it through a linguistic description. Therefore, the meaning of the symbol 'elephant' is extrinsic to the system. The symbol grounding problem may be solved by "interfacing a linguistic analyzer with a vision system." [MP96, p. 140]. However, how to connect them in an appropriate manner is an open question.

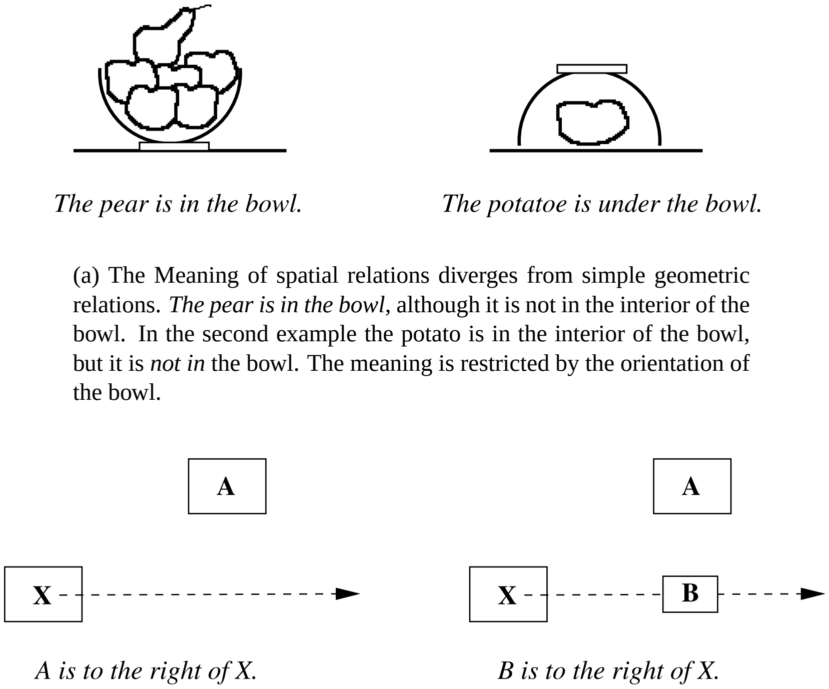
Connecting lexical word entries with visual recognizers would not be a sufficient solution. The referential competence does induce the grounding of the whole phrase. Therefore, it has to interpret the 'elephant', the 'living-room', *and* the preposition 'in' in combination.

2.4.3 Spatial cognition

The studying of such locative expressions is a research topic of its own and is often referred to as ***spatial cognition***. Indeed, there are psychological evidences that "appreciation of an object's qualities and of its spatial location depends on the processing of different kinds of visual information" [UM82, p. 578] and are localized in different specialized cortical areas. Ungerleider and Mishkin distinguish the ***"what"- and "where"-systems*** of the human brain. Nevertheless, locative prepositions cannot be interpreted independent of the involved object types and shapes, rather without considering context and background knowledge (see Fig. 2.5), as impressively shown by Annette Herskovits [Her86]. Barbara Landau and Ray Jackendoff argue that shape is used by the "where"-system in a very sparse and abstract sense whereas the "what"-system exploits shape information in a very detailed manner [LJ93]. They draw a parallel to the observation that we use more than 10.000 nouns to describe object classes but only about a hundred prepositions to describe object locations. Herskovits mentions two central questions of spatial cognition. "The ***decoding*** question is: given a locative expression used in a particular situation, can one predict what it conveys, how it will be interpreted – that is, provided it has been used approximately? If not, can one explain the inappropriateness? The ***encoding*** question is: given a situation with two spatial objects, can one predict the locative expression that can be used truly and approximately to describe their spatial relation?" [Her86, p. 11]. The answering of these questions results in a huge number of different computational models that try to formalize spatial relations under different constraints and contexts. Section 2.5.2 will discuss some of the relevant aspects of such spatial models.

2.4.4 A categorization of computational systems

The primary goal of computer science is not to discover principles, representations, and processes that humans might use for perception and cognition. Its primary goal is the design and construction of computer systems that realize a predescribed functionality. Therefore, a system which shall communicate with a human partner in a natural way must share some capabilities with humans such as understanding speech, recognizing objects in the environment, or interpreting locative expressions but need not implement these functions in the same way. Nevertheless, it is often useful to adapt principles discovered in cognitive psychology or linguistics because they describe a system which has the intended functionality – i.e. the human being.



(a) The Meaning of spatial relations diverges from simple geometric relations. *The pear is in the bowl*, although it is not in the interior of the bowl. In the second example the potato is in the interior of the bowl, but it is *not in* the bowl. The meaning is restricted by the orientation of the bowl.

(b) The applicability of the spatial relation *right* between *X* and *A* is influenced by the introduction of another contextual object *B*. Now, in many situations only *B is to the right of X*.

Figure 2.5: Divergences, unexplained restrictions, and unexpected context dependencies in meaning and use of locative expressions (from [Her86, p. 14,15,16]).

A good overview of computational models and applications for integrating linguistic and visual information is given by Rohini K. Srihari [Sri94]. She distinguishes between systems dealing with a single input stream, either language or visual inputs, that rely on integrated visual/language knowledge bases, and systems incorporating both linguistic and pictorial input streams.

A first representative of the first category is ***Natural Language Assisted Graphics*** (NLAG). “In such systems, a natural-language sentence is parsed and semantically interpreted, resulting in a picture depicting the information in the sentence.” [Sri94, p. 188]. Waltz proposes an “event simulation mechanism” for such purposes [Wal81]. It shall be capable of making plausible judgments about descriptions, is necessary for the resolution of anaphoric reference, circumvents short-term-memory limitations, and is a basis for mental imagery [Kos94]. Other work, mainly concentrates on the interpretation of locative expressions [ADG84, OMiT94] which establishes the understanding-modeling link to spatial cognition 2.4.3.

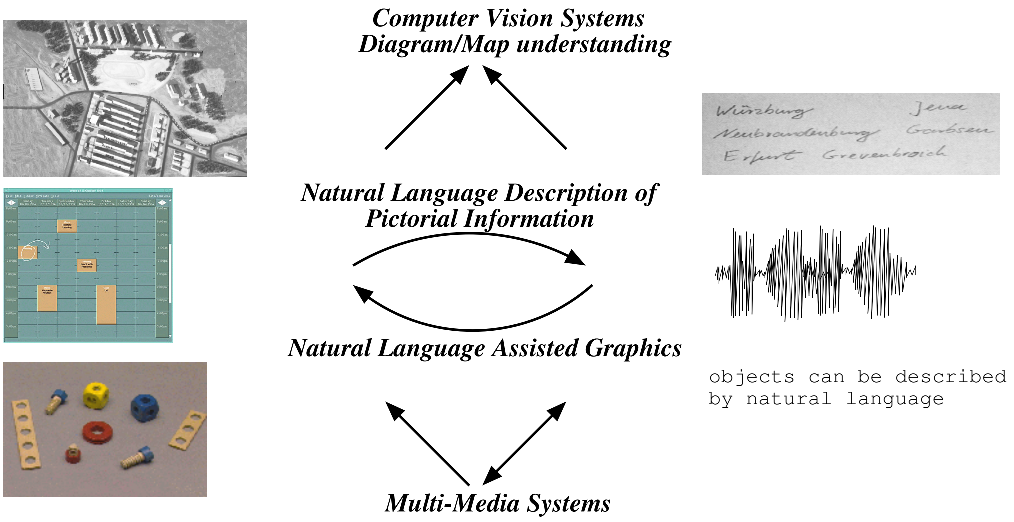

Computer Vision Systems
Diagram/Map understanding
Würzburg
Jena
Neubrandenburg
Garbsen
Erfurt
Grevenbroich
Natural Language Description of Pictorial Information
Natural Language Assisted Graphics
objects can be described by natural language
Multi-Media Systems

Figure 2.6: A categorization of computational systems.

The counterpart of NLAG is the generation of ***Natural-Language Descriptions of Pictorial Information***. “The problem is to generate a coherent text describing relevant objects, relationships between objects and events which are implicit in the output of the vision system.” [Sri94]. Typical domains are the descriptions of traffic scenes [NN83, Nag94, Nag99, HB00], the generation of sports commentary [HR95, LVW98], best path descriptions in landmark navigation [AK99], or the description of locations in medical images like radiographs [AK99]. ***Optical Character Recognition*** (OCR) may also be seen as a function that generates a symbolic description from pictorial input [Sri94]. However, the characteristics of such a process is quite different from those mentioned earlier. The input data is inherently symbolic as it is just another coding of text. Nevertheless, the recognition of handwritten text remains a challenging problem.

The second category includes systems that incorporate linguistic and pictorial inputs. Srihari classifies them into four distinct areas: (i) diagram understanding, (ii) map understanding, (iii) computer vision systems, and (iv) multimedia systems. ***Diagram understanding*** is the problem of producing an integrated meaning of combined groups of primitives (lines, curves, text, icons) which have to be extracted by a segmentation process (see Sec. 2.2). Thus, it is possible to interpret documents such as maps, weather maps, engineering drawings, business graphics, etc. [NB90, Raj94]. ***Map understanding*** is mentioned by Srihari as an extra subcategory which closely related to diagram understanding. She gives two examples: The system of Yokata et al. uses a common intermediate representation in order to present information given as visual input verbally and linguistic information pictorially in a weather report system [YTK84]. The system of Reiter and Mackworth uses a formal framework in order to interpret geographic maps [RM87]. Correspondences between domain, image, and scene knowledge are thereby specified in an explicit manner. In the context of incorporated linguistic and pictorial inputs, ***computer vision systems*** consider situations where pictures are accompanied by some descriptive text. The main idea of these systems is to benefit from the interpretation of text in image analysis and vice versa [AST81, TR87, SB94, ZV88].

The last area is that of ***multimedia systems*** (cf. e.g. [May93]). These "integrate data from various media (e.g., paper, electronic, audio, video) as well as various modalities (e.g., text, tables, diagrams, photographs) in order to present information more effectively to a user." [Sri94, p. 194]. Therefore, the design of intelligent user interfaces, which are able to automatically determine the referents when using deictic gestures and speech with reference to a visualized diagram or chart, is a main topic in this area. In contrast to computer vision systems, multimedia systems have total control of the common visual field which is internal instead of external to the system. This eliminates the need for an image interpretation system.

An additional area not mentioned by Srihari are ***audio-visual processing systems*** which benefit from a combined audio-visual input on a lower level than the linguistic processing level. They utilize the fact that human speech is bimodal both in production and perception when lip movements can be observed. That the human perception of speech is affected by the visual cue of lip movements has been shown by McGurk and MacDonald (see "McGurk Effect" [MM76]). These effects are used in applications such as audio-video coding, audio-visual speech recognition, or person verification [CR97].

The central question in all system categories mentioned is how to correlate audio/linguistic and image/graphical data. This is also known as the ***correspondence problem*** [Sri94, p. 350] which will be discussed in more detail in the next section.

2.5 The Correspondence Problem

Srihari defines the *correspondence problem* as "how to correlate visual information with words [...] [or more precisely with] events, phrases or entire sentences" [Sri94, p. 350]. Why should this correlation be difficult? Humans can do this very easily. The reason why human-to-human communication is very effective and robust is that both communication partners share common mental models of the world.

Visual information is a quantitative measurement of physical objects in the world. A speaker, who talks about the objects he or she sees, uses his subjective mental models in order to generate a description of the visual scene and to form the sentence he or she intends to utter (Fig. 2.7).

The communication partner has to perform a similar process. First, the message must be linguistically decoded in order to generate a semantic representation of the utterance. Secondly, it has to be referentially decoded, i.e. related to the physical world. The first process is based on common knowledge about used words, sentence structures, semantic meanings, i.e. linguistic knowledge. The second one is based on common mental models used in visual scene interpretation. The terms common knowledge and common mental models are written down easily, but human knowledge and human mental models are very difficult to represent explicitly and can only be implemented partially in a computer system. Even though linguistics and cognitive psychology have discovered many aspects

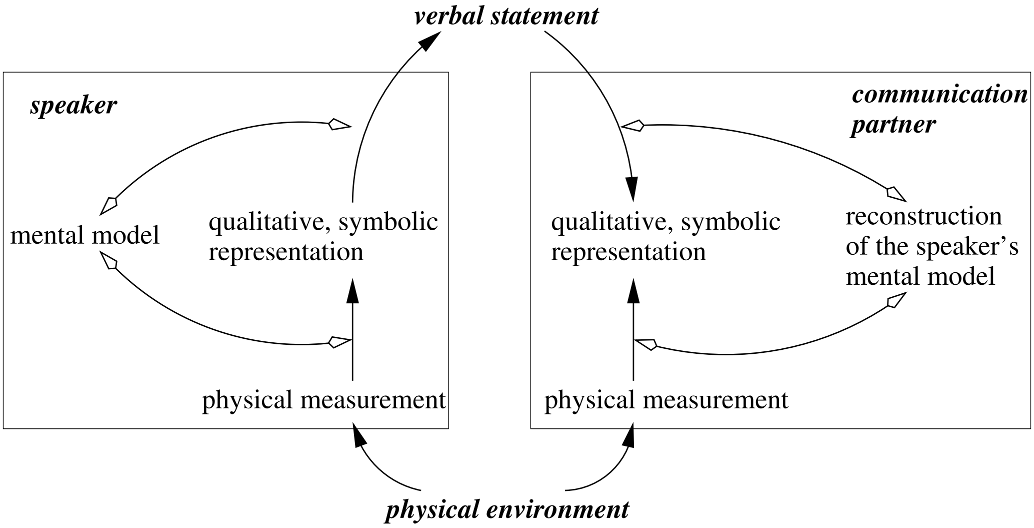

verbal statement
speaker
mental model
qualitative, symbolic representation
physical measurement
communication partner
qualitative, symbolic representation
reconstruction of the speaker's mental model
physical measurement
physical environment

Figure 2.7: The correspondence problem in human-computer communication: The 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.

of the human mind, the mental models, levels of processing and control strategies used by humans are not precisely known. In cognitive research, there are some computational models integrating speech and image processing that are motivated from the standpoint of psychology and linguistics (cf. e.g. the model of Jackendoff discussed in section 2.5.1). However, this thesis will take a more technical standpoint:

Postulate 1 *The correspondence problem is treated as an encoding/decoding process. Thus, we lose restrictions on the design of the computer system and can apply the proposed general integration approach even to technically motivated implementations in specialized domains.*

The input of the *complete decoding process* is the speech signal and an image or image sequence. The result of the decoding process is an explicit description of the speaker's intention. Note that the image interpretation task can be regarded as a partial decoding process of the image. Hence, the complete decoding process can be divided into three parts that are interrelated (see Fig. 2.8): (i) decoding of the speech signal (speech understanding) (ii) decoding of the image data (computer vision) (iii) decoding of the referential links.

In this encoding/decoding framework, *natural-language assisted graphics* consists of an encoding of the image description, i.e. generating the image, and thereby encodes the referential links. *Natural-language description of pictorial information* consists of a decoding of the image data and an encoding of the semantic description, i.e. generating natural language or speech, and thereby encodes the referential links (see Fig. 2.8).

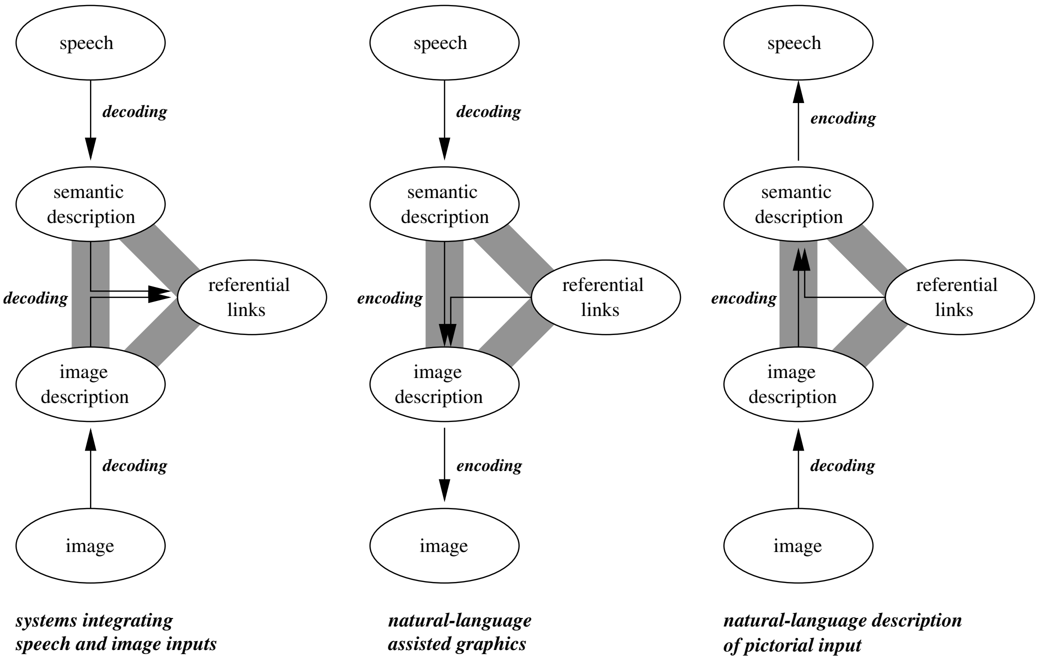


Figure 2.8: The correspondence problem in an encoding/decoding framework.

In this framework, the decoding and encoding of referential links can be identified as the central processing subtask of systems integrating speech and image processing. However, the border to the other subtasks is not always clearly defined. An extreme case of a de/encoder of referential links is a static integrated knowledge base, e.g. linking the lexicon entry for the word *'red'* with a specialized object recognizer for *red-objects*. Consequently, the referential coding process would be subsumed by the speech and vision decoding subtask. In the following sections it will be argued that a separate active inference process is much more flexible and that a referential coder should be treated as a separate subtask:

Postulate 2 *The referential de-/encoding process is organized as a separate subtask that is realized by an active inference process.*

The main problems of the encoding process lie in the fact that there may be thousands of different semantic descriptions of the same pictorial information and, on the other side, there may be an arbitrary number of pictures denoting the same semantic description. Instead of selecting an arbitrary encoding, the task has to fulfill diverse restrictions, like simplicity, specificy, typicality, etc.

The decoding process suffers from different problems. Ambiguity is closely related to the arbitrary number of corresponding representations in the encoder. The restrictions applied in the encoding step are not known. In the decoding process this has the consequence that there may be more than one valid decoding result or decoding results with different degrees of validity. Another problem is the occurrence of errors in the other subtasks, namely the decoding of speech (speech understanding) or decoding of the image (computer vision).

The solution of the referential coder problem has many different facets. What kind of knowledge representations are useful for such integration purposes? Do we need integrated knowledge bases, or can the integration process be managed by separate control structures and a universal inference calculus? Can the correlation function be learned? How should an integrated system be evaluated in order to show its robustness and efficiency? Besides these more computational aspects, the solution of the correspondence problem in a specific application is a question of modeling. The modeling task has to define the semantics of the concepts used in an application. On the one hand, the models used must be adequate to the application. On the other hand, they should be general enough in order to be able to switch to another domain. Much effort has been spent on the investigation of general purpose qualitative spatial models (cf. e.g. [Her86]). However, a framework that is general enough has not been discovered so far. Most computational spatial models (see e.g. [AK99, Gap94, MWH93, FSSS97]) or dynamic event models (see e.g. [Nag94, HB00, HKM$^+$94]) are restricted to domain-specific assumptions. Although some of them have been verified or calibrated on real data (see e.g. [VSF$^+$97, AK99]), models that are able to learn spatial or temporal relations from data are rarely reported. An example for learning dynamic event models is presented by Siskind [Sis98].

2.5.1 Knowledge representation and control structures

This thesis is not the first and not the only one that is dealing with the problem of how to relate visual and verbal information. This section will discuss three approaches which exemplify some principles in the fields of knowledge representation and control structures. Each of these approaches describes a different perspective on the same problem. However, there are many aspects they have in common.

The level of translation

Ray Jackendoff treats the integration of auditive and visual information from a psychological standpoint [Jac87, Jac89]. He aims at discovering general principles that can be applied to human cognition: “How can we talk about what we see?” [Jac87, p. 90]. The study of Jackendoff is based on the 3-d model of Marr and his own theory of Conceptual Semantics that will be briefly described below. It is argued that both theories are well suited for the formulation of translation rules between vision and language representations. He exemplifies this aspect by means of the notions of physical objects and spatial expressions.

The logical organization of language and vision that is used in the argumentation of Jackendoff is shown in Fig. 2.9. The organization of vision is based on the theory of David Marr and H. Keith Nishihara [Mar82]. It is composed of three different levels: (1) In the ***primal sketch*** the intensity image is converted into a representation that makes the locations of edges and other surface details explicit. It can be thought of a set of array cells that contain symbols indicating the presence of edges, corners, bars, and blobs of

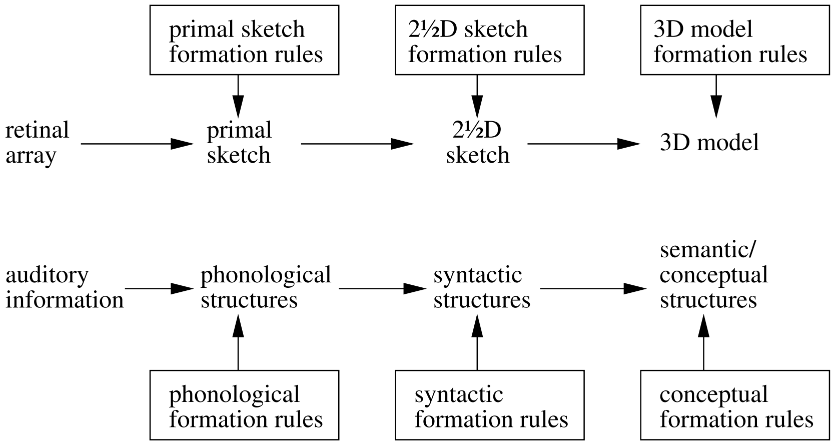

primal sketch formation rules
2½D sketch formation rules
3D model formation rules
retinal array
primal sketch
2½D sketch
3D model
auditory information
phonological structures
syntactic structures
semantic/ conceptual structures
phonological formation rules
syntactic formation rules
conceptual formation rules

Figure 2.9: Logical organization of language and vision [Jac87].

various sizes and orientations. This two dimensional representation is then transformed into (2) the $2\frac{1}{2}$***-d sketch*** by adding the third dimension using stereo, movement, shades, sizes of texture, etc. This representation consists of an array of cells that correspond to particular lines of sight from the viewer's vantage point. Each cell contains a set of symbols that define the depth and orientation of the local surface patch and indicate discontinuities in depth and orientation. The $2\frac{1}{2}$-d sketch is intended to comprise the richest possible information that early vision can deliver. (3) The next stage is the ***3-d model***. It is defined in object-centered, model-based, decompositional terms in contrast to the view-point specific, data-driven representations in the lower levels. Objects are represented in volumetric terms using ***generalized cones*** [Bin71]. The whole object is represented by a coarse shape description using a single generalized cone. Then the object is decomposed into its parts (*elaborated*) resulting in a finer shape description using a generalized cone for each part. These may be again elaborated into subparts, and so on. Any decomposition is represented in the coordinate system of the upper level. Variable shapes like a walking man may thereby be described very easily.

In language processing Jackendoff distinguishes three levels of representation: (1) The ***phonological structures*** describe the formation of words. (2) The ***syntactic structures*** combine syntactic categories (noun, verb, etc.) into phrasal categories (S, NP, etc.) considering aspects like case, gender, number, and tense. (3) The ***semantic/conceptual structures*** are based on a theory of semantics called ***Conceptual Semantics*** [Jac85]. The fundamental aspects of this theory can be characterized by the following four points: (a) *Meanings are mentally encoded.* Independent of the language user, truth statements are not related to "the world". They are justified in the reconstruction (or mental representation) of the world by the speaker. (b) *Meanings are decompositional.* Any syntactic structure of a sentence can be mapped to a sentential concept in the semantic theory. But sentential concepts cannot be listed and must be generated from a finite set of primitives and principles of combination. Even lexical concepts cannot consist of a list of instances. They must be build up from finite schemes that can be compared to novel inputs. (c)

Meanings do not, however, decompose into necessary and sufficient conditions. Concepts have fuzzy borderlines and bear resemblance to properties of other concepts. (d) *There is no formal distinction of level between semantics and pragmatics.* This proposes that even nonlinguistic tasks such as object categorization can be managed by the same principles of combination.

The only levels that include the notion of a physical object are the 3-d level of Marr and the conceptual level of Jackendoff. Therefore, these are the only levels possible for translation. An important correspondence that can be established between these two levels is the *part-whole* relation. It can be found in the decompositional meanings of Conceptual Semantics and is represented in the 3-d model in structural terms.

Jackendoff shows that Conceptual Semantics can easily be used to represent spatial expressions, like $[_{State}\mathrm{BE}([_{Thing}\mathrm{BOOK}], [_{Place}\mathrm{ON}([_{Thing}\mathrm{TABLE}])])]$, that contribute to the asymmetry between the reference object TABLE and the localized object BOOK. Such structures can easily be translated into Marr's 3-d model. The conceptual structure $[_{Place}\mathrm{ON}([_{Thing}\mathrm{TABLE}])]$ corresponds to a volumetric representation of a place that is bound to the volumetric representation of the TABLE. The state BE can be translated into a geometric test between the 3-d model of the BOOK and the 3-d model of the place.

Such correspondences can be found or can be easily defined for nearly any primitive concept or ***semantic part of speech*** in Conceptual Semantics, such as *Object, Place, Path, Action, State, Event*.

From text to visual constraints

Rohini Srihari et al. focus on the development of efficient control mechanisms for incorporating picture-specific context for image interpretation tasks in a newspaper domain [SB94, Sri95, CS95]. They use the interpretation of the text accompanying a picture in order to establish object hypotheses that are, then, localized and identified in the image. This is realized in a goal-driven top-down process that exploits a set of constraints that has been previously extracted from the text. Therefore, two aspects are addressed by Srihari: (1) How to represent and extract visual information from texts. This is solved by the definition of ***visual semantics***. (2) How to use this information in vision processing. This is realized by a constraint satisfaction technique.

The linguistic and visual semantics is organized in an integrated knowledge base that is realized in a KL/1 style semantic network formalism which is implemented in LOOM [SB94]. Each word in the lexicon is represented as a LOOM concept that is organized in a WordNet consisting of *is-a* and *has-part* hierarchies. Some of the entries are linked to *visual superconcepts* which reflect type (man-made, natural), shape, texture properties, boundary properties etc. of the object. *Visual is-a* and *visual has-part* links thereby superimpose a visual hierarchy on the WordNet concept hierarchy. Using this hierarchy, objects can be modeled on various resolution levels. A tree may be characterized as a natural object with a fractal boundary, may be described by its visual parts and their spatial relationships, or may be linked to a specialized recognition procedure for trees.

The visual semantics can be used to link words, phrases and sentences to visual information. For example, the knowledge base entry for a “hat” contains an inference rule stating that, if the hat is mentioned together with a human, it can be found above the head of the human.

Such *spatial constraints* are then used by the image understanding process. Besides this geometric or topological information, the NL module generates *characteristic constraints*, i.e. properties of objects like the sex or hair-color of a human, and *contextual constraints* that are e.g. predicted objects like the people present in a photo, or a classification of the general scene context like outdoor or indoor scene.

The vision control loop consists of the following three steps: (1) select a set of object classes of interest, (2) locate objects, (3) find a consistent labeling of the located objects. Each step may be influenced by the generated constraints. The selection of object classes can be set to 'human face' if the caption mentions a number of persons. The text may indicate the organization of the faces in rows that can be used in face location. In the last step, spatial constraints may be used to attach a name label to the faces. The process of satisfying various constraints, spatial and others, result in repeated calls to the image understanding module.

The system has been applied to a newspaper domain. The task was to locate and identify human faces in a photograph. The constraints were generated automatically from the captions of the images.

From visual primitives to verbal descriptions

Hans-Hellmut Nagel contributes to the problem of how to link conceptual descriptions in terms of natural language expressions to the results of image evaluation [Nag94, Nag99, HN00]. He identifies the system-internal conceptual representation as a "principle system component in its own right, independent of the 'surface modalities' between which it mediates" [Nag99, p. 80]. The task of the mediating component is an active one, namely the generation of the internal conceptual descriptions from modality primitives. Therefore, the representational structures are closely related with an inference machine – in Nagel's case fuzzy metric, temporal logic.

His application domain is the interpretation of traffic scenes. Driving vehicles like cars or busses are observed and their trajectories are characterized by conceptual descriptions that are closely related to natural language terms. This is based on the hypothesis that "natural language has evolved in order to cope with the complexity of everyday life" [Nag99, p. 81].

The interface between natural language descriptions and quantitative spatio-temporal patterns, i.e. the trajectories of the vehicles, consists of a set of primitive concepts and a terminology for building more complex concepts. The primitive concepts define the elementary vocabulary of the terminology to be used, e.g. $behind(vehicle,t,object)$, $enter(vehicle,t,area)$, $move(body,t)$, $with\_increasing\_speed(vehicle,t)$. The association of these concepts to the observed spatio-temporal patterns is realized by specific recognition procedures. The evaluation of each primitive predicate symbol yields a certainty value $\in [0,1]$.

Non-primitive concepts are constructed according to *specialization* and *(de)composition* rules. Decomposition refines a concept by specifying a subset of component concepts and relations between them. Specialization is realized by the conjunction of the given concept with an additional differentiating concept. The specialization/decomposition hierarchy can be elaborated to a *situation tree* that can be directly used for inferences. If a child situation can no longer be instantiated the search process reverts back to checking the parent situation. If the parent situation is consistent alternative children may be checked.

Nagel mentions a series of experiments regarding the expansion and scalability of the system-internal representation of knowledge about expected traffic situations. A systematic approach to the traffic domain identified 60-120 relevant motion verbs in the German language [HKN91]. Therefore, one main issue is the organizational structuring of the knowledge base. A graph-like representation of admissible action sequences had to be extended to hypergraph-like structures that introduce a hierarchical composition and generalization. The hypergraph structures were then rebuilt in fuzzy metric, temporal logic in order to facilitate a more controllable way to incrementally expand the knowledge base.

Principles of relating visual and textual information

Jackendoff claims the existence of a notion of an object as a prerequisite for representation levels that are suitable for integration of signal modalities. Such a notion exists in the two other frameworks as an agent that causes the trajectory (Nagel) or the object class that should be recognized (Srihari). However, the three-dimensional volumetric representation is no prerequisite in Srihari's visual semantics. The generated constraints can even applied on a simple visual blob level considering the blobs as unspecific objects or parts of an object. Thus, some aspects of a viewer-independent representation of objects are lost. Nevertheless, no principle problems are caused if the relation of the camera view and the reference frame of the (virtual) speaker is known, which is the case in the newspaper domain.

The structuring of the knowledge base by decomposition and specialization hierarchies is proposed by all three authors. But while Nagel is linking the quantitative visual information only to primitive concepts and formulates the hierarchy closely related to natural language terms, Srihari proposes linked but separated linguistic and visual hierarchies and a modeling of visual recognition on different levels depending on the precision of the textual information.

The treatment of spatial relation in Srihari's and Nagel's framework differs from that proposed by Jackendoff. Instead of modeling places as an entity of its own, they directly calculate relations between objects by evaluating fuzzified predicates or by applying constraints.

In the presentation of Conceptual Semantics Jackendoff states that meanings have fuzzy borderlines. This is reflected by Nagel in the usage of fuzzy logic. Within the framework of Srihari, characteristic constraints generate a thresholded confidence measure based on (i) how well the criteria are satisfied and (ii) the reliability of the routine itself. The confidence value is used afterwards to evaluate multiple solutions and to derive partial solutions.

In both frameworks, presented by Srihari and Nagel, a control mechanism for relating visual and verbal information is proposed. In Srihari's framework this component is merely controlled top-down by the textual information which is assumed to be correct. Possible actions of the vision component are directly linked in an *integrated knowledge base*, called visual semantics. However, bottom-up recognition is still possible. Due to a currently missing textual input in the system presented by Nagel, the natural language descriptions are calculated bottom-up. But a more flexible control seems also possible. He emphasizes that the system-internal representation that mediates between the visual and verbal modalities should be treated as an *independent system component*. Therefore, his approach does not propose an integrated knowledge base. Nevertheless, the terminology that is used to build up higher-level descriptions is closely related to language concepts. Jackendoff neither proposes an own control mechanism nor a separate representation level for integration purposes. Instead, he defines translation rules between the level of Conceptual Semantics and the 3-d model of Marr. Such a scheme is quite similar to the visual semantics of Srihari. Any link in the integrated knowledge base corresponds to a translation rule in the Jackendoff style.

2.5.2 Spatial models

Spatial models are a very important topic in systems that integrate verbal and visual information. Understanding a visual scene, understanding verbal descriptions, and establishing correspondences between objects mentioned and objects seen is strongly based on the spatial arrangement of objects in the scene. This section is merely based on an introductory article by Amitabha Mukerjee [Muk97], the book by Annette Herkovits [Her86], an article by Theo Herrmann [Her90], and partly the work of Daniel Hernández [Her94, CFH97].

The spatial arrangement of objects can be characterized by spatial relations between objects that correspond to linguistic expressions that may be used in a verbal statement, like *"the chair is in front of the table"*. Such spatial expressions partition the space around the table in a very loose fashion and with a large degree of ambiguity [Muk97, p. 1]. There are many positions around the table where a chair may be called *"in front of"*. However, if we compare two positions one is more likely to be called *"in front of"* than the other. The meaning of a spatial relation depends on inherent properties of the objects involved, like their relative distance, orientation, shapes, but also on the specific context in that a relation will be named. Spatial relations are typically classified into two basic categories. "***Topological relations*** are able to describe all aspects of the scene which are invariant with respect to common linear transformations (translation, rotation, rubber sheeting)" [CFH97, p. 319], for example *"the dog is in the kitchen"* or *"there are several chairs in a row"*. ***Projective relations***, Hernández calls them orientation relations, "describe where objects are placed relative to each other" [CFH97, p. 319], e.g. *"the fork should be placed to the left of the plate"*.

As discussed by Amitabha Mukerjee, there is a trend from ***neat*** approaches, using distinct, well defined symbols, to ***scruffy*** approaches, handling degree and context [Muk97]. The first approach employs a qualitative paradigm. The space is discretized into a set of qualitative zones or ***acceptance areas*** that are used to define appropriate predicates. An example for the one-dimensional space are the thirteen ordering relations defined by Allen [All83] based on an interval calculus. This approach is merely used for calculating topological relations like *in-contact* or *aligned*. It emerges to be very expressive in modeling relations in the zones near contact, e.g. *ON* [Muk97, p. 3]. Various mechanisms dealing with overlapping and non-overlapping acceptance areas have been discussed by Daniel Hernández [Her94]. The second approach models a gradation in a continuum. Either spatial relations are defined as fuzzy classes over the quantized space employing a continuous membership function (see e.g. [FSSS97]), or possible locations of objects are modeled by potential fields that can be tuned using a set of parameters (see e.g. [Gap94, OMiT94]). This approach is merely applied to non-contact or non-alignment positions like *near, far, in-front-of*.

Clementini et al. argue that "positions in space are likely to be represented in the [human] mind in a mixture of imaginal and propositional formats" [CFH97, p. 318]. Therefore, mixture models of qualitative and quantitative aspects might be promising.

When designing a spatial model for an application, we have to answer the question: what do we need to model? There are various aspects that may be relevant:

- **dimensionality:** shall the model work in the 1-d, 2-d or 3-d space? The relation of two cars on a road may be adequately described in one dimension. In order to describe a chair in front of a desk, a two-dimensional model is needed.

- **topology:** are objects in contact or aligned? There may be different definitions of contact relations based on overlap, touching, etc.

- **position and orientation:** how shall the relative pose between objects be described? Using discrete sets, continuum measures, constraint propagation? The selection may depend on the domain. In small-scale environments such as "the objects in a room", combinations of topological and orientational relations may be more relevant. For large-scale environments such as the geometric space, distance relations have to be considered [Her94]. The first scenario might be easily modeled using discrete sets, the second one by employing continuum measures.

- **scale:** typically, quantitative measurements must be related to the size of an object. Additionally, when interpreting the relation of objects further away from each other, it might be advantageous to change the level of abstraction. A bike that is parked near a big house might be more distant from the house than a tree that is near the bike. On the other side, the relation of the cities Bielefeld and Hanover will not be affected by the size of the cities because it is interpreted on a coarser scale, i.e. the relation between two points on a map.

- **shape:** the question of how the shapes of the involved objects influence the meaning of the spatial expression is a key issue in spatial modeling and remains an actual research topic in cognitive science. The possible positions of someone sitting in front of a table highly vary if the person is sitting at a long or short side of the table. Most approaches simplify the shape of an object by rectangular approximations or axis-based modeling.

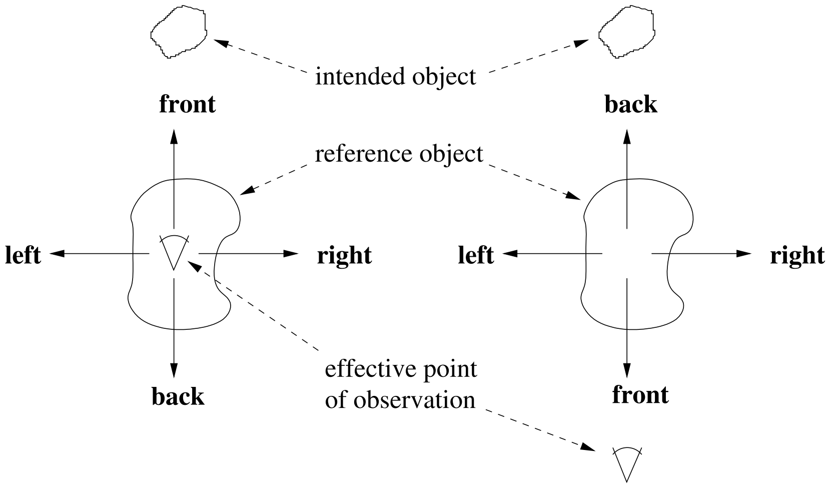


Figure 2.10: Projective relations are determined by the selection of the reference frame (left: basic ordering; right: mirror ordering) [Her86].

- **multiple objects:** most spatial relations are binary. Exceptions are for example the often used trinary relation *between* or patterns like *"place them in a circle"*. Nevertheless, even binary relations may be influenced by context objects. A chair will be called in front of a table only if there is nothing between them that introduces another context. If there is a bar between the table and the chair the bar gives the context of the chair, and it is not directly related to the table.

- **integrating time:** shall the spatial relations be applied to events that happen over time, e.g. one car overtaking another?

Another fundamental aspect must be handled for ***projective relations*** like *in-front-of*. Given an object arrangement, any choice of a projective relation automatically involves the selection of a ***reference frame*** that determines the directional meaning of a projective relation. There are different taxonomies which classify different selections of the reference frame. Hernández mentions three basic concepts that are needed to describe a projective relation: "the primary [or localized] object [subsequently, the term ***intended object***[2] will be used], the ***reference object*** that anchors the projective relation, and the ***frame of reference***" [CFH97, p. 319] (cf. Fig 2.10).

Clark introduces two fundamental cases, the basic ordering and the mirror ordering (see Fig. 2.10) [Cla73]. Retz-Schmidt distinguishes ***deictic***, ***intrinsic***, and ***extrinsic*** axes of the reference frame [RS88]. In the deictic case, directions are entirely defined by an observer, the speaker. Intrinsic directions are defined by inherent spatial properties of the object that anchors the relation, i.e. the reference object. The latter extrinsic case is closely related to the intrinsic case. External factors, e.g. motion, instead of inherent properties, impose a particular orientation on the reference object. Herskovits identifies

[2] The term ***intended object*** denotes the object that is specified by a verbal object description. Localizing this object by spatial relations is only one possibility of specification.

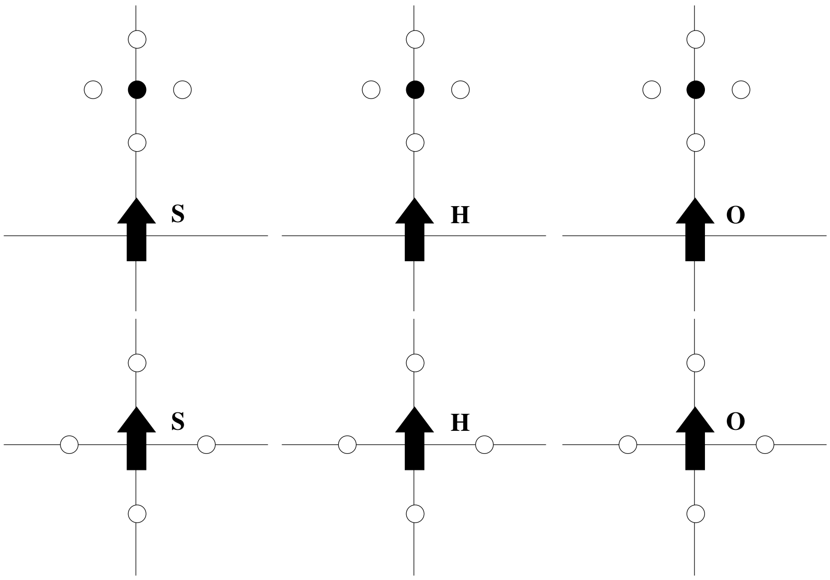


Figure 2.11: The six main variants of using projective relations as proposed by Theo Herrmann. The two rows distinguish three-point (view point, reference object, intended object) and two-point (view point and reference object are identical) localization. The columns distinguish the perspective used by the speaker (S), the hearer (H), or a context object (O) [Her90].

these two properties, mirror ordering and intrinsic/deictic, as independent of each other [Her86]. She mentions additional aspects that may contribute to a finer categorization of reference frame selections introducing two-point and three-point localization. These lead to a taxonomy as introduced by Theo Herrmann, that distinguishes six main variants of using the projective relations *front, back, left, right* (see Fig. 2.11) [Her90]. The three possibilities of two-point and three-point localization correspond to three different grammatical persons: *from my perspective, behind you, in front of the car*.

The design of a spatial model for a specific application can be simplified by limiting the vocabulary, using coarse shape descriptions and considering the *normal use* of a spatial relation. However, there are some aspects that are fundamental for the interpretation of locative expressions: they have gradual meanings, their meanings depend on context, and they are selected according to a specific reference frame.

2.5.3 Learning

As stated before, the correspondence problem in relating auditive and visual information is closely related to the symbol grounding problem in the language processing community. What is the meaning of a symbol? How can a system check the truth condition of a symbolic expression? How can it acquire the meaning of new symbols? The application of learning algorithms to correlated auditive and visual inputs might be a solution to these problems. There would be both theoretical and practical implications. On the one hand, such algorithms can provide models to test principles of how children perform language acquisition. On the other hand, the fixed vocabulary problem in human-computer interfaces may be circumvented. Verbal terms that are unknown to a system in the lexical or semantic sense may be adopted during a dialog.

Learning strategies in artificial intelligence, especially in the natural language processing community, can be divided into five different classes [Col94]:

- ***rote learning***: knowledge presented is duplicated by the learner.
- ***learning by instruction***: the knowledge is transformed into an internal representation used by the learner using a trivial quantity of preprocessing.
- ***learning by deduction***: the learner derives truth-preserving inferences from the knowledge available.
- ***learning by analogy***: existing knowledge is used to recognize similar situations. Knowledge from a previous problem is transferred to a new domain.
- ***inductive learning***: it is a similar process to that of deductive learning. But the truth preserving assumption may be violated by newly obtained knowledge.

This section will concentrate on ***inductive learning*** that can be also interpreted as reconstructing a function from a set of input/output examples, namely the mapping from words to meanings. Inductive learning algorithms differ in the representation used to describe the goal function and the feedback that is available (see e.g. [RN95, chap. 18]):

- ***supervised learning***: inputs and outputs of the function can be perceived.
- ***reinforcement learning***: the system receives some evaluation of its decisions but is not told what decisions were correct.
- ***unsupervised learning***: the system does not get any hint about the correctness of a decision.

In the case of learning a lexicon from correlated auditive and visual signals, input and output information is partly available but typically noisy and uncertain due to multi-word utterances, multiple contexts and recognition errors. Jeffrey Mark Siskind counts five reasons why such a lexical acquisition task is very difficult indeed [Sis96]:

- **multi-word utterances:** which words in an utterance map to which parts of the utterance meaning?
- **multiple contexts:** which of the context objects/events is in fact the meaning of the utterance just heard?
- **start without prior knowledge (bootstrapping problem):** how do children start the lexical acquisition process without any seed information?

- **input is noisy:** the corpus used for learning may contain utterances only paired with incorrect hypothesized meanings. Which input should be ignored?

- **many words are homonymous (can have several different senses):** which sense of each word is used at a given time?

The approaches proposed in the literature differ in the level of preprocessing assumed for the input. This section will discuss two different systems. The system of Deb Roy and Alex Pentland [RP98b, RSP99] is able to learn the correspondence of auditive and visual information on a very low, i.e. near to the signal, representation level involving the learning of visual concepts and new words. Correspondences are established on the lowest levels proposed by Jackendoff: the primal sketch and phonological structures. Siskind [Sis96] assumes that a speech recognizer provides a sequence of uttered words and a vision component is able to provide qualitative descriptions of visual events that are translated into Jackendoff-style conceptual expressions. In [Sis98] he gives some ideas how such a visual event recognizer may be learned in a supervised way. The language acquisition task is then to learn a lexicon from sequences of words paired with sets of possible conceptual expressions describing the visual context.

Learning low level correspondences

Deb Roy and Alex Pentland describe a system that incorporates four types of learning: (i) visual concept learning, (ii) learning new words, (iii) learning simple syntactical word ordering, (iv) learning the correspondence between visual concepts and words.

The aim is to learn an audio-visual lexicon from correlated noisy acoustic input and color images. The acoustic input may consist of natural multi-word utterances. Therefore, the word learning task includes the segmentation of the acoustic input into words. The figure background segmentation on the vision side is simplified by using a uniform background and by avoiding occlusions.

In the training phase the correspondence problem is solved by pointing on the referenced object or presenting a single object to the system. A phoneme recognizer that consists of an all-phoneme loop hidden Markov model (HMM) and a phoneme transition bigram calculates the most likely phoneme trace from the acoustic input. It achieves a phoneme recognition accuracy of about 70%. The visually observed objects are separated from the background and are characterized by a color and a shape histogram. The combined phoneme trace and histograms of the presented object are subsequently called acoustic-visual events (AV-events). First, a sufficient number of AV-events is accumulated. For each AV-event several word hypotheses are extracted by variable splitting of the phoneme trace. The word-object pairs are reduced by several filter criteria, like prosodic highlight, recurrency of speech segments, etc.

In order to find final word-shape or word-color clusters that constitute new words and their visual categorical meanings, Roy and Pentland introduce separate distance measures between visual events and between acoustic events that are combined using a mutual information measure.

The distance between two speech segments a, b is defined on the basis of a probability measurement. The phoneme recognizer calculates the most likely phoneme sequences $Q_{a/b}$ of the segments a resp. b. From these specific HMMs $\lambda_{a/b}$ are generated using the phonemes as states and connecting them in a strictly left to right manner. State transition probabilities are inherited from context-independent phoneme models. The distance is based on the "cross"-production probabilities, that is the probability that Q_a is produced by the HMM λ_b and vice versa:

$$d_A(a,b) = -\frac{1}{2}\left\{\log\left[\frac{P(Q_a|\lambda_b)}{P(Q_a|\lambda_a)}\right] + \left[\frac{P(Q_b|\lambda_a)}{P(Q_b|\lambda_b)}\right]\right\}$$

where $Q_{a/b}$ is the phoneme sequence of a resp. b and
$\lambda_{a/b}$ is the HMM derived from the speech segment a resp. b.

The distance of two visual events is measured by the χ^2 divergence of the associated histograms X, Y:

$$d_V(X,Y) = \chi^2(X,Y) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$$

The calculation of the mutual information measure of a word-shape pair $\mathcal{X}$ depends on two variable thresholds defining a cluster around the audio-visual event $\mathcal{X}$. Two variables $A, V \in \{0,1\}$ indicate the resulting membership of other word-shape hypotheses of the cluster of X. The two thresholds are optimized using the maximum mutual information (MMI) $I(A;V)$ as a criterion:

$$I(A;V) = \sum_{s\in\{0,1\}} \sum_{t\in\{0,1\}} P(A=s, V=t) \log\left[\frac{P(A=s,V=t)}{P(A=s)P(V=t)}\right]$$

The measure is high if the two events, that a word-shape pair $\mathcal{Y}$ is an element of the auditive interval ($A = 1$) and that it is an element of the visual interval ($V = 1$), are highly correlated. Thus, the mutual information measures a distance between the two probability distributions $P(A = s, V = t)$ and $P(A = s)P(V = t)$. The selection of the final word-shape clusters is performed using a greedy strategy. Successively, the hypotheses with the highest MMI is selected and all other hypotheses which match an optimized cluster both visually and acoustically are deleted. In a final step remaining clusters are selected according to a threshold applied to the mutual information score of the cluster.

In experiments this learning strategy turned out to be very robust and effective. Its most powerful characteristic is the generic representation of visual objects and words. No previous modeling and no manual adaptation to new domains is needed. In [RP98a] Roy and Pentland even present a first step towards syntax learning in that they generate a co-occurrence statistics of the acoustic entries in the audio-visual lexicon that is used in speech recognition. However, the aim of a boot strapping speech and image understanding system is quite far away.

Learning high level correspondences

Mark Jeffrey Siskind presents a more structured approach to learning the visual meanings of words. He assumes as a prerequisite that a speech recognizer provides a sequence of words for each auditive input and that a vision system is able to produce a conceptual description in the Jackendoff style from the visual scene input, e.g. the sentence *"John walked to school."* while seeing the visual event GO(John,TO(school)). While hearing the spoken utterance, the learner may see several events happening simultaneously, each of which would be a possible meaning of the utterance. The aim of the system is to learn the correspondence of such sentences and conceptual expressions. Siskind refers to it as the ***mapping problem***. It can be divided into two stages, that are realized in an interleaved manner: (i) The system learns the set of conceptual symbols used to construct the conceptual expression that corresponds to a word, e.g. *"raise": { CAUSE,GO,UP }*. (ii) The system learns how to compose these conceptual symbols, e.g. *"raise": CAUSE(x,GO(y,UP))*.

The result of the learning process is a word lexicon that is consistent with all pairings of utterances and possible meanings. Therefore, an utterance meaning must be broken down into parts and correctly assigned to the individual words. The task is complicated by multi-word utterances, multiple contexts, noisy input, homonymous words, and by starting without prior knowledge.

Siskind realizes his learning algorithm by exploiting four common sense principles that have been previously proposed by various psychologic researchers (cf. e.g. [Pin89, FHRG94]). Siskind refers to it as ***cross-situation learning***:

1. **constraining hypotheses with partial knowledge:** The system has previously learned that a word must refer to a conceptual symbol or does not refer to a conceptual symbol. This knowledge can be applied to the possible meanings of a new utterance in order to reduce the set of possible meanings.

2. **cross-situation inference:** The system finds something common across all observed uses of a word. Thus, possible meanings of words can be reduced.

3. **covering constraints:** All components of the meaning of an utterance must be derived from the meanings of words in that utterance. If a meaning fragment of the conceptual expression of the utterance is ruled out for all words of an utterance except for one, this fragment must correspond to the remaining word.

4. **principle of exclusivity:** Words in an utterance meaning must contribute to non-overlapping portions of the utterance meaning. If a meaning fragment is a necessary meaning part of one word in the sentence it cannot be part of the meaning of another word in the sentence.

From these principles Siskind has formulated inference rules (Fig. 2.12,2.13) that are applied to the actual representation of words in the lexicon. This representation consists of three different sets defining the possibly uncertain meaning of a word *w*:

Rule 1 *Ignore those utterance meanings that contain a conceptual symbol that is not a member of $\mathcal{P}(w)$ for some word symbol w in the utterance. Also ignore those that are missing a conceptual symbol that is a member of $\mathcal{N}(w)$ for some word symbol w in the utterance.*

Rule 2 *For each word symbol w in the utterance, remove from $\mathcal{P}(w)$ any conceptual symbols that do not appear in some remaining utterance meaning.*

Rule 3 *For each word symbol w in the utterance, add to $\mathcal{N}(w)$ any conceptual symbols that appear in every remaining utterance meaning but that are missing from $\mathcal{P}(w')$ for every other word symbol w' in the utterance.*

Rule 4 *For each word symbol w in the utterance, remove from $\mathcal{P}(w)$ any conceptual symbols that appear only once in every remaining utterance meaning, if they are in $\mathcal{N}(w')$ for some other word symbol w' in the utterance.*

Figure 2.12: Inference rules for learning conceptual symbol sets [Sis96].

- $\mathcal{P}(w)$ is the set of possibly corresponding conceptual symbols. It can be interpreted as an upper bound of the correct meaning. Initially, it consists of all conceptual symbols in the domain.
- $\mathcal{N}(w)$ is the set of necessarily corresponding conceptual symbols. This can be interpreted as a lower bound of the correct meaning. It is initially the empty set.
- $\mathcal{D}(w)$ is the set of possibly corresponding conceptual expressions. Initially, it comprises all allowed combinations of the possible meaning symbols.

By processing the set of utterances paired with hypothesized meanings, the representations of the words in the lexicon converge successively to the correct meanings. For example, let us assume that the following lexicon has been learned (example from [Sis96]):

	$\mathcal{N}$	$\mathcal{P}$
John	{John}	{John, ball}
took	{CAUSE}	{CAUSE, WANT, GO, TO, arm}
the	{}	{WANT, arm}
ball	{ball}	{ball, arm}

Now suppose that the algorithm receives the utterance

"John took the ball."

with the following hypothesized meanings:

(1) CAUSE(John,GO(ball,TO(John))
(2) WANT(John,ball)
(3) CAUSE(John,GO(PART-OF(LEFT(arm),John),TO(ball)))

Rule 5 *Let RECONSTRUCT$(m, \mathcal{N}(w))$ be the set of all conceptual expressions that unify with m, or with some subexpression of m, and that contain precisely the set $\mathcal{N}(w)$ of non-variable conceptual symbols. For each word symbol w in the utterance that has converged on its actual conceptual-symbol set, remove from $\mathcal{D}(w)$ any conceptual expressions not contained in RECONSTRUCT$(m, \mathcal{N}(w))$, for some remaining utterance meaning m.*

Rule 6 *If all word symbols in the utterance have converged on their actual conceptual symbol sets, for each word symbol w in the utterance, remove from $\mathcal{D}(w)$ any conceptual expressions t, for which there do not exist possible conceptual expressions for the other word symbols in the utterance that can be given as input to COMPOSE along with t to yield one of the remaining utterance meanings as its output.*

Figure 2.13: Inference rules for learning conceptual expressions [Sis96]. The output of COMPOSE is the set of conceptual descriptions that denote possible ways of combining the given word sense meanings into utterance meanings.

Rule 1 eliminates the meanings (2) and (3). Rule 2 eliminates the possible meanings 'arm', and 'WANT' from the lexical entries of *took*, *the*, and *ball*. Rule 3 adds the symbols 'GO' and 'TO' to the necessary meanings of *took*. Rule 4 eliminates the symbol 'ball' from the possible symbol set of *John*, yielding the following lexicon:

	$\mathcal{N}$	$\mathcal{P}$
John	{John}	{John}
took	{CAUSE, GO, TO}	{CAUSE, GO, TO}
the	{}	{}
ball	{ball}	{ball}

Rule 5 and 6 are concerned with the possible conceptual expression that can be composed from the necessary conceptual symbol sets. By applying rule 5 the algorithm converges on the conceptual expressions of the words *John, the* and *ball*, but leaves two possibilities for the word *took*: CAUSE(x,GO(y,TO(z))) and CAUSE(x,GO(y,TO(x))). Rule 6 eliminates the possible expression CAUSE(x,GO(y,TO(z)) from the set $\mathcal{D}$(*took*) because there is no word left that may provide a possible meaning that may be unified with z.

The algorithm is applied in an on-line and single pass way so that it can process any new utterance without considering previous utterances a second time.

The strategy described so far will fail if words have multiple meanings, for example in case of *homonymy* or noisy utterances, i.e. no hypothesized meaning corresponds to the utterance. As a consequence, a lexical entry might become *corrupted*: Either an impossible conceptual symbol is added to the necessary set or a necessary symbol is removed from the possible set. Both cases cannot be directly detected because the correct lexicon is not known. Instead, Siskind uses a weaker criterion in order to detect corrupted lexical entries. An entry is called *inconsistent* if one of two invariants is violated: either the necessary set remains no subset of the possible set or the possible set is empty. An inconsistent lexical entry is necessarily corrupted, but not the inverse. In such a case, the representation of a word is split into two senses of the same word. Senses that are not supported by enough evidence from the corpus are treated as noise. The two phenomena homonymy and noise are thereby captured and dealt with by the same algorithmic strategy.

Learning versus modeling

Both studies discussed in this section intend to simulate language acquisition strategies in early childhood. In the case of human-computer interfaces, the situation differs in that one does not need to start from scratch. Learning strategies need not *build up* the knowledge base, they shall *expand* the knowledge base. Consequently, learned items in the knowledge base must be compatible with modeled items. Additionally, they shall expand it in an *incremental way*. Therefore, new evidential items must be processed one after the other resulting in a new learning state after each item.

In order to apply Deb Roy's approach in such an environment, some prerequisites must be fulfilled. The distance measures used in the auditive and visual domains must be applicable to modeled items. The cluster generation needs an explicit training phase and a sufficient number of training examples. Therefore, it is difficult to apply during a dialog in an online way.

Siskind represents the ambiguity of word meanings in a more explicit way by enumerating possible meanings. On the one hand, this results in very large representations if the meaning cannot be constrained by other items in the knowledge base. On the other hand, existing modeled items can just be employed for this purpose. The representations of learned and modeled items are automatically compatible because the same qualitative representations produced by the vision component are used. The learning scheme can be used and is applied by Siskind in an incremental way tracking all possible alternatives. Therefore, it could be easily applied in a dialogue situation. The drawback of the learning scheme proposed by Siskind is that all *unknown words* must be known by the speech recognizer and all objects and visual events must be modeled in the vision component.

2.6 Other Related Work

Computer vision systems that incorporate pictorial and verbal information have been developed in many application areas. In the following, some of them that were influential or seem to be promising will briefly be reviewed concentrating on the kind of integration they propose and the application they realized it for.

One of the earliest systems integrating pictorial and verbal information is that of Abe, Soga and Tsuji [AST81]. They describe a system that understands the plot of a story by referring to both a series of line drawings with colors and narrations in English concerning this drawings. The vision part of the system is realized in a top-down fashion using structured object models. The conceptual description of the story is constructed employing a rule-based approach. After that, the system can answer questions about the story.

Lazarescu et al. use natural language understanding and image processing to index and query a database of American football tapes [LVW98]. Spatio-temporal characteristics are automatically extracted and represented by Allen's temporal primitives. They extract the type of the happening action, the players involved in the action and some game statistics like the score of the game from the commentary text. One task, incorporating this information, is the labeling of the detected player positions from the video. The system is able to find plays that are similar to a query play.

A probabilistic framework that fuses video and audio information for semantic indexing is presented by Naphade et al. [NKFH98]. They show two examples of detecting explosions and waterfalls in a video database. First, the audio and the video tracks are processed separately by hidden Markov models (HMMs). A supervisor HMM that encodes the correlation of states in both modalities then fuses the optimal state sequences found by the Viterbi algorithm in the separate video and audio HMMs. The result of the supervisor HMM is the detection of a probabilistic multi-media object (Multiject), e.g. the occurrence of an explosion. High-level probabilistic dependencies between the different possible multijects are organized in a graphical network (multinet). For example, the detection of the multiject "Beach" will increase the probability of the detection of "Yacht" or "Sunset" and decrease the probability of "Snow Clad Mountains".

Many approaches for video indexing are applied to news videos, e.g. [IHTS98, IT98, SSS$^+$97]. These image sequences are highly structured by topics, have simple settings, and include a single news speaker with a good pronuncation.

Takahashi et al. present a multi-modal user interface for a robot [TNKS98]. The robot can be instructed by speech and gestures, e.g. "Bring that apple" while pointing on it. In order to remove ambiguities in robot tasks, the control strategy includes the possibility to ask the user for further information. They use a frame-based production system. The aim of the system is to fill the object slots by collecting information from every available cue.

SAM (speech activated manipulator) is a robot system that interacts with a human instructor [BBW92]. The vocabulary of the speech recognizer comprises about 200 words and is designed in relation to the technical capabilities of the system. Objects are localized by an ultra-sonic range finder that is located in the gripper of the robot arm. In a training phase the robot learns to characterize the object by its position, color, and general shape and stores a verbal description of the object for later reference.

PLAYBOT is a long term project that aims at the developement of a prototype environment which will assist disabled children in play [TVD$^+$97]. The hardware platform consists of a stereo camera head, a robot arm for grasping, and an ActiveDeskTop that is a large-scale, touch-sensitive video display. The main aspect of the research is to use vision as the primary sensor of the system that short-circuits the control loop between the instructing children and the robot arm. The system design is based on a behavior based architecture. Each behavior either performs actions on external physical objects or on internal (logical) representations. Visual behaviors include visual attention, gaze stabilization, object recognition, object tracking, object search, event perception, calibration, and hand-eye coordination. Non-visual behaviors include the processing of the PLAYBOT command language and the object grasp behavior. The language parsing and semantic analysis component reads the sequence of touches on the ActiveDeskTop and translates them into well-formed commands for the robot. The object recognition component is based on the detection of geons, small volumetric parts, that may be combined in order to represent more complex objects.

Another application domain is augmented reality. The idea is to enrich the real world with an electronic information space that can be used to provide further descriptions of objects, e.g. in art galleries, to facilitate navigation in certain places, or other explanatory information. Nagao and Rekimoto present an *Ubiquitous Talker* that recognizes real world objects by scanning attached bar codes [NR95]. It classifies the current situation with regard to a situation library that employs a non-linguistic context to the speech understanding part of the system. Based on the situation awareness, the speech recognizer is constraint by selecting the vocabulary and grammar for analyzing the spoken utterances. The user can verbally select an item from the displayed menu on a palmtop computer or ask questions about the displayed information.

The system of Bronsted et al. utilizes a frame-based integration scheme that is realized in a blackboard architecture [BLM+98]. They have developed a multi-media workbench which can be used as a campus information system. A blueprint of a building layout is placed on the workbench table and queries can be formulated by speech or by pointing with a stick.

Another category of system are multi-media systems that concentrate on error correcting strategies using speech/NL and pictorial inputs. Waibel et al. present a system incorporating speech, gestures, handwriting and face tracking as input modalities [WSVY97, SMW96]. They develop diverse strategies to explicitly correct previously given input, like respeaking, spelling, repair by handwriting, selecting among N-best, or using pen gestures. They give a measurement based on accuracy and needed time in order to predict the strategy the user will select. Their approach has been tested in a medical application called QuickDoc, which helps a doctor to quickly identify, label, and comment anomalous areas in a series of images such as X-rays or computer-aided tomography scans.

In summary, the variability of applications for an integrated processing of visual and verbal information is vastly increasing. However, most of the systems simplify the processing of one of the input channels by using active displays, range finders, or bar codes. Only a few of the approaches examine how different noisy channels combine. Waibel et al. develope correcting strategies for multi-modal intefaces, Naphade et al. define a super HMM in order to robustly combine the the input data streams. Nearly all systems are limited to a dedicated domain. Only the PLAYBOT vision system realizes a first step to the recognition of arbitrary shaped objects. The frame-based slot filling interactions scheme is very popular. However, most approaches do not take account of contradictory slot contents.

2.7 Contributions

This work is a new solution of the correspondence problem in correlating speech and images. It will be applied in a ***computer vision system*** (cf. Sec.2.4.4) that incorporates both speech and pictorial data. Both input channels are analyzed by specialized vision and speech understanding components that perform partial decoding processes on the input signals. The task of establishing referential links between the partial decoding results is treated as a third decoding process. The result of the decoding process is a set of possible assignments of the verbal description of scene objects and their visual representation.

2.7.1 A probabilistic translation scheme

The modeling of this decoding task must contribute to different kinds of uncertainty in order to make the solution robust despite of noise, propagated errors, and vague meanings. Therefore, translation rules must be probabilistic in contrast to logical rules or links in a knowledge base. Translation rules must be modeled on different abstraction levels, like Srihari's blob level and Marr's composed 3-d objects, recognized words and structured object descriptions. This thesis will show that a unified modeling is still possible if any partial result of a vision or speech understanding component is interpreted as an evidence in a probabilistic network. In order to exploit the redundancy in the verbal and visual description of an object, two different kinds of information must be integrated:

- Information about the object class or category.
- Information about the spatial and structural context of the individual object in the scene.

These two aspects can be found in all three approaches of Jackendoff, Srihari, and Nagel. However, only the scheme proposed by Srihari exploits spatial constraints for the object labeling process in order to increase robustness.

The probabilistic network used for integration partially reconstructs the mental models of the speaker. For this purpose, a mixture of explicit and implicit modeling is proposed. One part of this reconstruction is reflected in the structure of the network, another part is reflected in the probability numbers that can be estimated using the corpus of the application domain or calculated from simulation models.

2.7.2 A separate integration and interaction component for speech understanding and vision base-line systems

Today, a computational system that performs vision in an universal and confidential manner, just as proposed by David Marr, does not exist. A computational system that understands spontaneous free speech in an universal way independent of any domain or environmental condition does not exist either. Any existing computational system is in some way specialized to the domain it is realized for and the paradigm it is realized with. A translation scheme that is based on such universal processing schemes like that of Jackendoff currently seems not feasible. Therefore, an integration component must in some way stand for its own independent of the techniques used for analyzing the surface modalities. This aspect is best worked out in the approach of Nagel who proposes an independent hierarchy and inference calculus for integration purposes. A drawback of Nagel's approach is the fact that the interface between the specialized vision modules and the integration component, i.e. the primitive concepts, is very small. This thesis will broaden this interface by a more detailed modeling of the vision component.

The base-line system consists of a speech understanding component that is able to extract simple instructions and (partial) object descriptions from spontaneous speech (see Sec. 4.3.1). The vision component extracts a finite set of elementary objects using a feature-based approach and structural knowledge. Furthermore, the structure of composed objects is analysed in a generic way (see Sec. 4.3.2).

If referential links between both representations are established the same probabilistic network that is used to solve of the correspondence problem can now be utilized in order to draw inferences between both modalities. Some examples will be shown for the indexing step in object categorization, the verification of object hypotheses, and for the disambiguation of alternative verbal interpretations.

2.7.3 The choice of the application area

The application of the proposed integration scheme is a human-computer interface for instructing a robot that is able to take, assemble, or put parts on a table. On the one hand, this is a very simple domain. The elementary objects are known, the lighting conditions and background can be controlled, some object shape categories or types are correlated with a finite set of colors, the way how elementary objects can be joined together is known. The structure of spoken instructions is simple, spatial relations mostly refer to directions on the plane of the table.

On the other hand, complexity is introduced by two aspects. First, the technical names of the elementary objects are not known to the speaker. Shape descriptions like *"long"*, *"big"*, *"thin"* have gradations that depend on context. Secondly, complex objects that are constructed from elementary objects introduce occlusions, new shapes, and more complex object descriptions.

All these aspects can be controlled very well in this domain, which makes it a very good test domain for speech and image integration systems. The test domain will be described in more detail in section 4.1.

2.7.4 Inference and learning

The identification of verbally referred objects and the drawing of inferences between modalities are prerequisits of learning because they establish new facts about a particular situation. This thesis will show that the probabilistic integration component is able to link visual objects to unknown object names that were introduced by the speaker. By this means, a rudimentary semantics is assigned to the unknown names.

This task is even more difficult if instead of the complete assembly a subassembly is denoted by a speaker. In this case the boundary of the named part must additionally be learned. In an outlook section 5.3 a solution is presented that applies ***cross-situation inferences*** – similar to those of Siskind (see Sec. 2.5.3) – to a *sequence* of dialog steps.

Chapter 3

A Model for Uncertainty

Any reasoning task in a realistic domain requires simplifications. Conclusions are drawn although many facts about a situation are not available. Decisions are taken without considering all possibilities. Actions are performed before all consequences have been checked. Reasoning with uncertainties is something normal and trivial in our daily life, but something difficult to be exactly specified for a computer. Imagine the following situation:

Example 1 *You visit the Wimbledon Tennis Championships and walk beside the small court number fifteen, where a dark-skinned man with rasta-curls is playing against a white European with dark, extremely short hair. On the scoreboard you can see the names of the two players, 'Agenor vs. Lendl' but you do not know who is who. And you think about a strategy to find out ...*

The first idea might be to check for correspondences between names of the players and how they look like. If one name sounds French, for example, we would apply the inference rule *if someone has a French name then he will also look French.* However, before one can apply this rule, first, the conditions have to be checked under which this rule is *allowed* to be applied. In artificial intelligence this question is known as the ***qualification problem*** [McC77]. There might be several exceptions that cannot be enumerated completely, like "he might be a citizen of a former French colony and, therefore, has an African, Asian, or Polynesian look", "he might be a Brazilian who has been adopted by a French family in early childhood", "he might have a German looking parent from Elsass that is a French region with much German ethnic and cultural influences", etc. Therefore, we can never be sure that the inferred statement is true. Judea Pearl compares reasoning with exceptions with the navigation task in a minefield: "Most steps are save, but some can be devastating." [Pea88, p. 1]. An alternative way of enumerating all exceptions is to summarize them. It is like setting up some warning signs in the minefield to indicate that a specific area is more dangerous than others [Pea88, p. 1].

The treatment of such uncertainty measures is different to that of truth values. Pearl argues that two logic formulas $A \to C, B \to C$ can be syntactically combined yielding the truth value of $(A \wedge B) \to C$. However, it is not clear how exceptions should combine. “Whereas truth values in logic characterize the formulas under discussion, uncertainty measures characterize invisible facts, i.e. exceptions not covered by the formulas.” [Pea88, p. 2]. Consequently, the principle of modularity cannot be transferred to the uncertainty calculus, unless restrictive independencies can be explicitly assumed.

3.1 Intensional and extensional models

Pearl distinguishes two principle approaches to uncertainty treatment. ***Extensional models*** are a generalization of rule-based production systems (well known examples are MYCIN [Sho76], or PROSPECTOR [DHN76]). Any rule is attached with an uncertainty measure that is treated like generalized truth values. If the name in the given example sounds French with a measurement of $p = 0.9$ and we have the given rule

$$\textit{name sounds French} \rightarrow^{0.7} \textit{looks French},$$

the truth value that he will have a French look will be syntactically combined yielding e.g.: $p' = 0.9 \cdot 0.7 = 0.63$. The measurement $p' = 0.63$ summarizes the past inference process. If we now get a second information from the scoreboard that the player with the French name comes from Haiti and we have the rule

$$\textit{Haitian citizen} \rightarrow^{0.2} \textit{looks French},$$

the new information has to be combined with $p' = 0.63$ without considering that this information has been inferred from the French name, e.g. applying a rule from MYCIN:

$$p'' = 0.63 + (1.0 \cdot 0.2) - 0.63 \cdot (1.0 \cdot 0.2) = 0.704$$

The two facts *Haitian citizen* and *name sounds French* are treated as irrelevant to each other when deducing *looks French*.

Intensional systems, Jensen calls them ***normative systems*** [Jen96], do not model the inference process of the expert. They declaratively model the domain. The uncertainty measure is not coupled with rules, but attached to the *state of affairs* or subset of possible worlds [Pea88, p. 3]. In the subset of possible worlds, where the predicate *name sounds French* is true, the probability that the same person *looks French* is $P(\textit{looks French}|\textit{name sounds French}) = 0.7$. This is a statement about the domain, not about the inference process. Given this statement, we cannot infer anything until we know that the current situation is an element of the same subset of possible worlds. Starting with the possible world in that the name is Agenor we have to consider both possible worlds in that *’the name sound French’* is true or false before deducing something about a French look:

$$\begin{aligned}
&P(\textit{looks French}|\textit{name is Agenor})\\
&=P(\textit{looks French}|\textit{name sounds French})P(\textit{name sounds French}|\textit{name is Agenor})\\
&+P(\textit{looks French}|\neg\textit{name sounds French})P(\neg\textit{name sounds French}|\textit{name is Agenor})\\
&=0.7 \cdot 0.9 + 0.1 \cdot 0.1 = 0.64
\end{aligned}$$

The second information that the man is a *Haitian citizen* changes the subset of possible worlds. Therefore, we have to calculate:

$$\begin{aligned}&P(\mathit{looks\ French}|\mathit{name\ is\ Agenor},\mathit{Haitian\ citizen})\\ =&P(\mathit{looks\ French}|\mathit{name\ sounds\ French},\mathit{Haitian\ citizen})\\ &P(\mathit{name\ sounds\ French}|\mathit{name\ is\ Agenor},\mathit{Haitian\ citizen})\\ +&P(\mathit{looks\ French}|\neg\mathit{name\ sounds\ French},\mathit{Haitian\ citizen})\\ &P(\neg\mathit{name\ sounds\ French}|\mathit{name\ is\ Agenor},\mathit{Haitian\ citizen})\end{aligned}$$

Although, if the information that he is a *Haitian citizen* does not change the probability that the *name sounds French*,

$$\begin{aligned}&P(\mathit{name\ sounds\ French}|\mathit{name\ is\ Agenor},\mathit{Haitian\ citizen})\\ =&P(\mathit{name\ sounds\ French}|\mathit{name\ is\ Agenor}) = 0.9,\end{aligned}$$

it might reduce the conditional probability of *looks French* because in Haiti there are other ethnic influences, e.g.:

$$\begin{aligned}&P(\mathit{looks\ French}|\mathit{name\ sounds\ French},\mathit{Haitian\ citizen}) = 0.3\\ &\neq P(\mathit{looks\ French}|\mathit{name\ sounds\ French}) = 0.7.\end{aligned}$$

Therefore, the recalculated probability considering the information about the *Haitian citizen* yields:

$$P(\mathit{looks\ French}|\mathit{name\ is\ Agenor},\mathit{Haitian\ citizen}) = 0.3\cdot 0.9 + 0.1\cdot 0.1 = 0.28$$

In turns out that the independence assumption of the two facts *Haitian citizen* and *name sounds French* in the extensional calculation results in a different conclusion about the expectation of a *French look*. The point here is not that the combination rule of the extensional approach has to be changed in order to fix the outcome of the calculation. The point is that the knowledge implicitly coded in the combination rules of the external approach can be explicitly specified in the intentional approach. The price to pay, however, is an increasing number of parameters (conditional probabilities) that have to be specified as well as an increasing number of operations (multiplications, additions) that have to be performed.

Another aspect of intensional systems is their ability of bidirectional inference. Instead of inferring the visual appearance of the player from the given name, the same model can be used to infer the kind of name from the visual appearance, here by using the Bayesian rule from probability theory:

$$\begin{aligned}&P(\mathit{name\ sounds\ French}|\mathit{looks\ French})\\ &= P(\mathit{looks\ French}|\mathit{name\ sounds\ French})\frac{P(\mathit{name\ sounds\ French})}{P(\mathit{looks\ French})}\end{aligned}$$

For rule-based extensional systems the introduction of bidirectional rules would lead to circular inferences.

Without explicitly saying it, the intensional calculus applied in the previous example is exactly that of ***Bayesian networks***. It is a rather intuitive formalism that is well founded in probability theory. The language provided by Bayesian networks is very well suited for modeling probabilistic causal and relevance relations and will be discussed in more detail in the next subsections.

Besides Bayesian networks there are other uncertainty calculi that can be applied in an intensional way, like the ***Dempster-Shafer calculus*** [LGS86]. This chapter will concentrate on Bayesian networks, which will be applied throughout this thesis because of their foundation in probability theory, their capability to explicitly model relevance relations, and their possibility to apply bidirectional inferences. The last property is especially important in integrated processing of different modalities. The integration model must be able to draw inferences from – in this case – speech understanding to object recognition and vice versa.

3.2 Bayesian Networks

Bayesian networks are one possibility of modeling declarative knowledge in a realistic domain. Although Bayesian networks are numerically exact and mathematically well founded in probability theory, the basic concepts used for modeling correspond to relationships that are also useful in normal discourse (see [Pea88, p. 16]):

- **likelihood:** *$\mathcal{A}$ is more likely than $\mathcal{B}$*. This statement may be true either without any prerequisite or it may be true due to some observation O; or written formally:

  $$L(\mathcal{A}|O) > L(\mathcal{B}|O).$$

  In the example, it is more likely that the black man with rasta-curls has a French name than the man who looks like an Eastern-European. In Bayesian networks this can be formalized by interpreting $\mathcal{A}, \mathcal{B}$ as two different states of a random variable H and setting two probabilities

  $$P(O|H = \mathcal{A}) = p_1, P(O|H = \mathcal{B}) = p_2, \text{ with } p_1 > p_2.$$

- **conditioning:** *Given that what I know is C ....* This can be syntactically captured by placing C behind the conditioning bar in a statement like $P(A|C) = p$. p denotes the ***belief*** in some statement A. The notion $P(A|C)$ is also called ***Bayes conditionalization***. The definition of $P(A|C)$ is given by the famous ratio formula of Thomas Bayes:

  $$P(A|C) = P(A,C)/P(C) \tag{3.1}$$

  The belief that the man will look French is conditionalized by the information about the French name. *’Looks French’* and *’name sounds French’* are random variables with possible values $\{\text{true}, \text{false}\}$:

$$P(\textit{looks French}|\textit{name sounds French}) = \frac{P(\textit{looks French}, \textit{name sounds French})}{P(\textit{name sounds French})}$$

- **relevance:** *A and B are relevant to each other in context C*, if the likelihood of A would change if B is added to C:

  $$L(A|C) = P(C|A) \neq L(A|C,B) = P(C|A,B).$$

  The calculation of the belief in C under different contexts $P(C|A)$ and $P(C|A,B)$ captures exactly what is expressed by the term ***non-monotonic reasoning*** in artificial intelligence. The inferred knowledge about the French looking of someone who has a French name must be completely revised when the information is added that he is an Haitian citizen. In the other case, the Haitian citizen is not relevant if we infer that the name *'Agenor'* sounds French. The likelihood that the name sounds French does not change.

- **causation:** *A causes B* is a very intuitive notion in human reasoning. The probabilistic interpretation is based solely on the notion of relevance. ***Causation*** is a very strong tool for structuring and specifying probabilistic knowledge. B is a *direct cause* of A if the relevance relation between A and B is not affected by any other context information. Therefore, the notion of *A directly causes B* depends on the level of modeling detail. Thus, in a reduced model the name *'Agenor'* may directly cause a French looking. Introducing the predicate about the French name separates the French looking from the name *'Agenor'* that, now, causes it in a transitive sense.

  More complex relevance relationships can be specified by combining causations. If we assume that two reasons A, C to independently cause an event B, once we have observed the event the two causes become related. Thus, confirming A might lower the belief in C and vice versa. The alternative reasons are ***explained away***.

  Example 2 *If Ronald Agenor scores a point against Ivan Lendl, he either scores the point because of a well prepared attack or because of an unforced error of Lendl. Observing that the last stroke of Ronald Agenor was a well timed volley stop nearly rules out the assumption of an unforced error of Ivan Lendl.*

The perspective on probabilistic modeling, that is reflected in these basic modeling concepts, is best characterized by Glenn Shafer: “Probability is not really about numbers; it is about the structure of reasoning” [Pea88, p. 15].

Probabilistic reasoning is often criticized by statements like “How to get those numbers?” or “Why should beliefs combine like frequencies?”. Very often it is very difficult indeed to get exact estimates of all conditional probabilities needed. Sometimes, they must even be guessed from introspection. Pearl gives two reasons why probability calculus is, nevertheless, a good framework. “If we strongly believe in the rules by which exact quantities combine, we can use the same combination rules on the rough estimates at hand." [Pea88, p. 21]. "Second, when we commit ourselves to a particular set of numbers, no matter how erroneous, the consistency of the model prevents us from reaching inconsistent conclusions." [Pea88, p. 21].

3.2.1 Definition of Bayesian networks

A ***Bayesian network*** $\mathcal{B} = (\mathcal{V}, \mathcal{E}, \mathcal{C})$ is a graphical representation of a joint probability distribution $P(U) = P(A_1, \dots, A_n)$ with:

- A set of variables $\mathcal{V} = \{A_1, \dots, A_n\}$ each having a finite set of mutually exclusive states $A_i \in \{a_1^{(i)} \dots a_{m_i}^{(i)}\}$.
- A set of directed edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$.
- The variables together with the directed edges form an directed acyclic graph (DAG), i.e. there is no directed path $A_i \rightarrow \cdots \rightarrow A_j$, with $A_i, \dots, A_j \in \mathcal{V}$ such that $A_i = A_j$. The set of nodes that have an edge pointing to the same child node A are called ***parents*** of A: $pa(A) = \{B_i | B_i \in \mathcal{V} \wedge (B_i, A) \in \mathcal{E}\}, A \in \mathcal{V}$
- A set of ***conditional probability tables*** (CPTs):
  $\mathcal{C} = \{P(A|B_1, \dots, B_n) \mid pa(A) = \{B_1, \dots, B_n\}, A, B_1, \dots, B_n \in \mathcal{V}\}$
  such that to each variable A with parents $B_1, \dots, B_n$ there is attached a CPT $P(A|B_1, \dots, B_n)$.
- The joint probability distribution $P(U)$ is given by the product of all CPTs: $P(U) = \prod_{P(A|B_1,\dots,B_n) \in \mathcal{C}} P(A|B_1, \dots, B_n)$

The notation in the following sections uses uppercase letters (e.g. A) for random variables and lowercase letters (e.g. a) for values of variables. $P(a)$ is the shortened version of $P(A = a)$. $P(A)$ denotes a ***probability table*** that comprises the probability values for any possible assignment of $A \in \{a_1, \dots, a_m\}$:

$$P(A) = \left[P(A = a_i);\ i = 1 \dots m\right] = \left[P(A = a_1), P(A = a_2), \dots, P(A = a_m)\right]$$

Bayesian networks provide an algebra for conditional probability tables. A ***value assignment*** $P(B|A = a_1, C)$ selects a subtable of $P(B|A, C)$:

$$P(B|A = a_1, C) = \begin{bmatrix} P(b_1|a_1, c_1), & P(b_1|a_1, c_2), & \dots & P(b_1|a_1, c_n) \\ P(b_2|a_1, c_1), & \dots & & \\ \dots & & & P(b_m|a_1, c_n) \end{bmatrix}$$

The ***multiplication*** of CPTs $P(B, C|A, D) = P(B|A, C)\ P(C|D)$ is defined as follows:

$$P(B, C|A, D) = \left[P(b_i|a_k, c_j)\ P(c_j|d_l);\ i = 1 \dots m, j = 1 \dots n, k = 1 \dots r, l = 1 \dots s\right]$$

Consequently, an observation of ***evidence*** $\mathbf{e} = \{A = a_2\}$ can be represented as a probability table by assigning 1.0 to the second component and 0.0 to the remaining components:

$$\underline{e} = \left[0.0, 1.0, 0.0, \dots, 0.0\right]$$

Multiplying by a CPT, e.g. $P(A)$, results in:

$$P(A,\mathbf{e}) = P(A)\ \underline{e} = \left[0.0, P(a_2), 0.0, \ldots, 0.0\right]$$

The ***division*** of CPTs $P(B|A,C,D) = P(B,C|A,D)/P(C|D)$ is defined analogically to the multiplication:

$$P(B|A,C,D) = \left[P(b_i,c_j|a_k,d_l)/P(c_j|d_l);\ i = 1\ldots m, j = 1\ldots n, k = 1\ldots r, l = 1\ldots s\right]$$

Normalizing a CPT $P(A,B=b_1,C,D=d_2|F,G=g_1)$ corresponds to a conditionalization of the probability distribution with regard to the fixed variables, i.e. the variables that have been assigned a value, e.g.

$$P(A,C|F,B=b_1,D=d_2,G=g_1) = \frac{P(A,B=b_1,C,D=d_2|F,G=g_1)}{P(B=b_1,D=d_2)}$$

Marginalization of a CPT $P(A,B|C,D)$ means that a variable is eliminated from the CPT by summation, e.g.

$$P(A|C,D) = \sum_b P(A,B=b|C,D) = \left[\textstyle\sum_b P(a_i,b|c_j,d_k);\ i = 1\ldots m, j = 1\ldots n, k = 1\ldots r\right]$$

Maximization is an other possibility for variable elimination, e.g.

$$\begin{aligned} P(A|C,D) &= \max_b P(A,B=b|C,D) \\ &= \left[\max\nolimits_b P(a_i,b|c_j,d_k);\ i = 1\ldots m, j = 1\ldots n, k = 1\ldots r\right] \\ \underline{B}(A,C,D) &= [\operatorname*{argmax}_b P(a_i,b|c_j,d_k);\ i = 1\ldots m, j = 1\ldots n, k = 1\ldots r] \end{aligned}$$

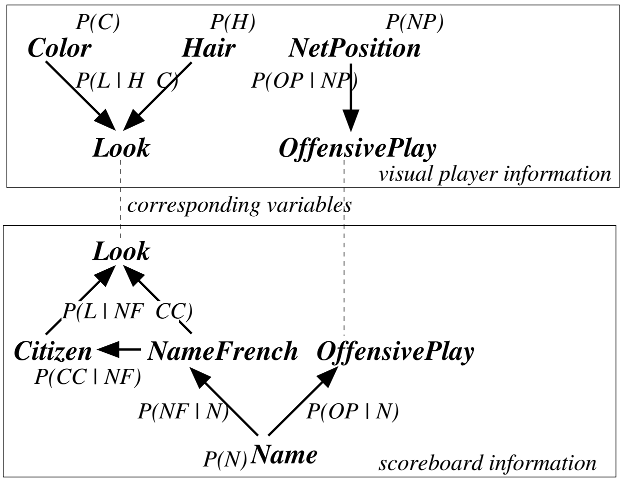
The matrix $\underline{B}(A,C,D)$ stores the maximized values of the operation variable, here B.

Marginalization and maximization can only be performed over variables that are *not* conditioned (here only A or B).

3.2.2 Modeling in Bayesian networks

In order to demonstrate how different kinds of information can be translated into the language of Bayesian networks, the story of Ronald Agenor and Ivan Lendl will be continued:

Example 3 (continued Ex. 1) *: You visit the court number fifteen at the Wimbledon Tennis Championships. On the scoreboard you can see the two names 'Agenor' from Haiti and 'Lendl' from USA. You do not know who is who. You think that 'Agenor' sounds like a French name, but no one on the tennis court looks like a Frenchman. However, one of the players is dark-skinned and has black rasta-curls which might be compatible with the French name and the country Haiti. Then you remember that you have read in the newspaper about a young offensive player named 'Agenor'. This information may fit to this dark-skinned player because he frequently takes a net position.*



Name (N)	∈ {*Agenor, Lendl*}	– name on scoreboard
NameFrench (NF)	∈ {*true, false*}	– name sounds French
Look (L)	∈ {*French, Eastern-European, exotic*}	– looks like
Citizen (CC)	∈ {*Haiti, USA*}	– citizen of country
OffensivePlay (OP)	∈ {*true, false*}	– offensive playing
NetPosition (NP)	∈ {*true, false*}	– net position
Color (C)	∈ {*black, white*}	– skin color
Hair (H)	∈ {*rasta-curls, other*}	– hairstyle

Figure 3.1: Example of two Bayesian networks in the tennis domain that model different kinds of information (cf. Ex. 3).

The story tells us something about the entries on the scoreboard and gives information about the players on the tennis court. Certainly, these two different kinds of information must be related, but how they correspond to each other is not known. Therefore, these will initially be modeled in separate Bayesian networks (see Fig. 3.1).

The first Bayesian network (scoreboard information) represents the joint probability distribution $P(L,OP,CC,NF,N)$. It models coherences between the look (L), an offensive play (OP), the citizenship (CC), the sounding of the name (NF), and the name of the player (N). Without restrictions the joint probability distribution can be written as the following product (applying the Bayes' conditioning Eq. 3.1):

$$\begin{aligned}P(L,OP,CC,NF,N) =& P(L|OP,CC,NF,N)\; P(OP|CC,NF,N)\\ & P(CC|NF,N)\; P(NF|N)\; P(N)\end{aligned}$$

Modeling in Bayesian networks means deciding about relevance relations. The first assumption is that the citizenship CC and the sounding of the name NF are direct causes of the look L of the player. Given these information, the offensive play OP and the exact name N are irrelevant:

$$P(L|OP,CC,NF,N) = P(L|CC,NF)$$

Secondly, the newspaper information did not include any detail about the country and we assume that there is no correlation between a name that sounds French and an offensive playing. Therefore, the offensive player OP is only directly related to the name of the player N:

$$P(OP|CC,NF,N) = P(OP|N)$$

Thirdly, the exact name of a player is not relevant to the citizenship of the player if the more abstract information of how the name sounds like is given. On the other side, the knowledge about the sound of the name will change our expectation about the citizenship:

$$P(CC|NF,N) = P(CC|NF)$$

Given these conditional independency assumptions, the joint probability distribution can be rewritten as:

$$P(L,OP,CC,NF,N) = P(L|CC,NF)\ P(OP|N)\ P(CC|NF)\ P(NF|N)\ P(N)$$

This is exactly what is represented by the Bayesian network *'scoreboard information'* in Fig. 3.1.

The second Bayesian network (*'visual player information'*) models coherences between the look of the player (L), his skin color (C), his hairstyle (H), an observed net position (NP), and an offensive playing (OP):

$$\begin{aligned} P(L,C,H,OP,NP) =& P(L|C,H,OP,NP)\ P(C|H,OP,NP)\ P(H|OP,NP) \\ & P(OP|NP)\ P(NP) \end{aligned}$$

The net position and offensive playing is assumed to be irrelevant to the look of the player. The skin color C and the hairstyle H are only modeled to be correlated if the information about a French, Eastern-European, or exotic look is given:

$$P(L,C,H,OP,NP) = P(L|C,H)\ P(C)\ P(H)\ P(OP|NP)\ P(NP)$$

The resulting Bayesian network structure is given in Fig. 3.1 (visual player information).

Before the modeling of the corresponding variables in the example will be described, the next section will focus on another problem in Bayesian networks: *How to get those numbers* – i.e. the conditional probability tables?

3.2.3 How to get those numbers? Some simplification

The design task of a Bayesian network consists of selecting an appropriate structure and determining the numbers of the conditional probability tables (CPTs). In networks that include nodes with many parents these CPTs can be quite large. In the following some modeling techniques will be described that simplify this task. Finn V. Jensen calls them “modeling tricks” (see [Jen96, pp. 47]).

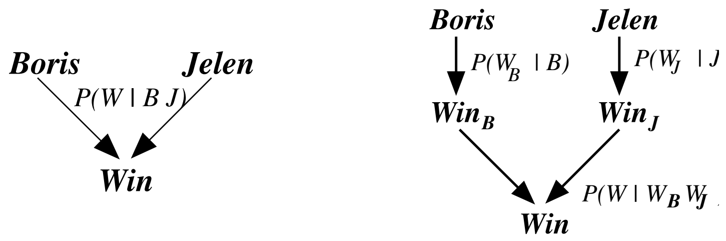


Figure 3.2: Noisy-or gate (cf. Ex.5).

Undirected relations

In many domains one has to model logic or probabilistic relations between random variables that do not have a direction such as causal relations.

Example 4 *You are still at the Wimbledon Tennis Championships, but now you are visiting a double with four players. On the scoreboard the four names of the players are paired in order to form the two teams: Becker/Jelen vs. Jarryd/Edberg.*

The pairing of the names defines an undirected binary relation *Team*. Let

$$\begin{aligned} &A,B \in \mathcal{S} = \{\text{Jarryd, Edberg, Becker, Jelen}\}, T \in \{0,1\}, \\ &\mathit{Team} = \{(\text{Jarryd,Edberg}), (\text{Becker,Jelen})\} \subset \mathcal{S} \times \mathcal{S}. \end{aligned}$$

The numbers of the probability table can be directly obtained from the definition of the relation:

$$\begin{aligned} P(T=1|A,B) &= \begin{cases} 1.0, & \text{if } (A,B) \in \mathit{Team} \\ 0.0, & \text{otherwise} \end{cases} \\ P(T=0|A,B) &= 1.0 - P(T=1|A,B) \end{aligned}$$

Noisy-or

A special case of an undirected relation is the ***noisy-or*** gate. If a probabilistic relation $P(A|B,C)$ is difficult to specify, but if it was possible to estimate $P(A|B)$ and $P(A|C)$, the conditional probability table $P(A|B,C)$ can be constructed from the simpler CPTs.

Example 5 (continued Ex. 4) *You intend to estimate the probability that the team (Becker,Jelen) scores the next point. They will score either because of a good playing of Becker, a good playing of Jelen, a good playing of both, or because of a fault of the opponents.*

The variable $W = \mathit{true}$ denotes that Becker and Jelen will score the point, $B = \mathit{true}$ that Boris Becker plays well, and $J = \mathit{true}$ that Erik Jelen plays well. The conditional probabilities $P(W|B)$ and $P(W|J)$ have been estimated. Now the assumption is introduced that

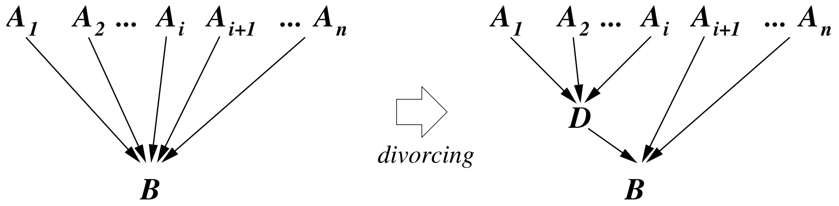

A_1 A_2 ... A_i A_{i+1} ... A_n
B
divorcing
A_1 A_2 ... A_i A_{i+1} ... A_n
D
B

Figure 3.3: The parents of node B are divorced by a new mediating node D.

the reasons B, J are independent of each other and combine like the logical OR, yielding:

$$\begin{aligned} P(W|B,J) &= \sum_{W_B,W_J} P(W|W_B,W_J)P(W_B|B)P(W_J|J) \\ \text{where } P(W|W_B,W_J) &= OR(W_B,W_J), \\ P(W_B|B) &= P(W|B), \\ P(W_J|J) &= P(W|J). \end{aligned}$$

Evaluating the OR relation yields that the noisy-or can be specified by three values q_0, q_1, q_2:[1]

$$\begin{aligned} P(W|\neg B,\neg J) =& 1 - P(\neg W_B|\neg B)P(\neg W_J|\neg J) = 1 - q_0 \\ P(W|B,\neg J) =& 1 - P(\neg W_B|B)P(\neg W_J|\neg J) = 1 - q_0 q_1 \\ P(W|\neg B,J) =& 1 - P(\neg W_B|\neg B)P(\neg W_J|J) = 1 - q_0 q_2 \\ P(W|B,J) =& 1 - P(\neg W_B|B)P(\neg W_J|J) = 1 - q_0 q_1 q_2 \end{aligned}$$

Divorcing

The noisy-or is a special case of a technique that is called ***divorcing*** [Jen96, p. 52]. The set of parents of a node B: $pa(B) = \{A_1 \dots A_n\}$ is divorced by a ***mediating variable*** D (Fig. 3.3). Divorcing by D means that a subset of parents $\{A_1 \dots A_i\}$ is substituted by the variable D:

$$pa(B) = \{D, A_{i+1} \dots A_n\}$$

such that D in turn becomes a child of the substituted set: $pa(D) = \{A_1 \dots A_i\}$. This operation is valid if $D \in \{d_1, \dots, d_m\}$ defines a corresponding partitioning $\mathcal{D}_1, \dots, \mathcal{D}_m$ of the set of configurations:

$$A_1 \times A_2 \cdots \times A_i = \bigcup_j \mathcal{D}_j \text{ with } \mathcal{D}_j \subseteq A_1 \times A_2 \cdots \times A_i,$$

so that if $(a_1, \dots, a_i), (a'_1, \dots, a'_i) \in \mathcal{D}_j$, then

$$P(B|a_1, \dots, a_i, A_{i+1}, \dots, A_n) = P(B|a'_1, \dots, a'_i, A_{i+1}, \dots, A_n)$$

[1] Note that the last probability can be calculated from the other probabilities: $(q_0 q_1 q_2) = \frac{(q_0 q_1)(q_0 q_2)}{q_0}$.

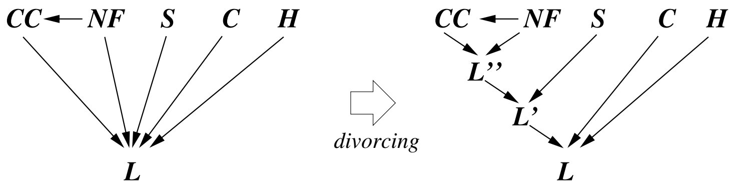


Figure 3.4: An example of divorcing (cf. Ex. 3, Fig. 3.1).

For example, if the evidence on the scoreboard and that from the visual appearance of the players on the tennis court shall be integrated to obtain a common belief in the look of the player (cf. Ex. 3, Fig. 3.1), divorcing can be used in order to reduce the size of the CPTs. The look of the player is modeled by the country CC and a variable NF indicating whether the name sounds French on the one side as well the skin color C and the hairstyle H on the other side. The mapping from the entry on the scoreboard to the player on the court is represented by a selection variable S which will be explained in the next section. In this case it will be sufficient to assume that the variable S controls the causal influence of the scoreboard entry on the look of the player on the court. Therefore, the probability distribution may be modeled as follows (Fig. 3.4, left):

$$P(L,C,H,CC,NF) = P(L|C,H,S,CC,NF)\,P(C)\,P(H)\,P(CC|NF)\,P(NF)$$

Divorcing the parents of node $L \in \Omega_L = \{French, Eastern\text{-}European, exotic\}$ by a new mediating variable $L' \in \Omega_L$ results in a modified structure of the Bayesian network

$$P(L|C,H,S,CC,NF) = \sum_{l\in\Omega_L} P(L|L'=l,C,H)\,P(L'=l|CC,NF,S)$$

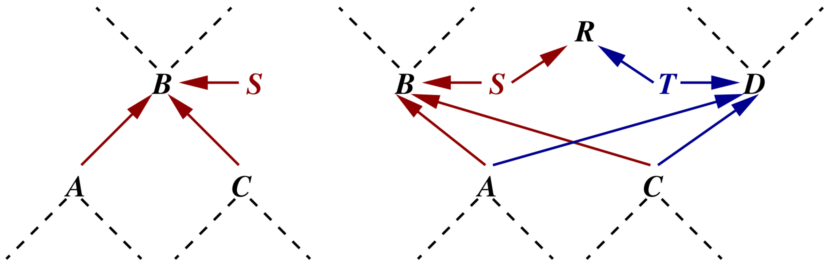
Here the visual evidences *'hair'* (H), *'skin color'* (C) and the scoreboard evidences *'citizen of country'* (CC), *name sounds French* (NF) are assumed to independently cause the impression of the look L. In a second step, the divorced subset of variables $\{CC,NF,S\}$ can be divorced again resulting in the structure shown in (Fig. 3.4, right):

$$P(L'|CC,NF,S) = \sum_{l\in\Omega_L} P(L'|L''=l,S)\,P(L''=l|CC,NF)$$

The selection variable S only considers the causal influence of CC and NF summarized in L''.

3.2.4 Modeling corresponding variables

The previous subsection discussed the Bayesian network structure (Fig. 3.1) of the tennis example (Ex. 3). The resulting structure of the network consisted of two subnetworks that were related by two ***corresponding variables***. We cannot link them directly because the first network models one of two entries on the scoreboard, and the other network models the look of one of the two players. There are four sets of evidences: $\{$*'black','rasta-curls'*$\}$, $\{$*'white','short-hair'*$\}$, $\{$*'Agenor','Haiti'*$\}$, $\{$*'Lendl','USA'*$\}$. We do not know if



(a) One-to-two mapping

(b) Two-to-two mapping

Figure 3.5: Modeling of corresponding variables.

Ronald Agenor or Ivan Lendl should be related to the dark-skinned man with rasta-curls or to the white man with short hair. Certainly, the question *'who is who on the tennis court?'* may be answered externally to the Bayesian network by setting up a search on every possible combination of evidence with some matching criterion. However, there is an internal modeling alternative:

Postulate 3 *The modeling of corresponding variables is a key issue in relating multimodal input. This thesis will show that this problem can be solved in the language of Bayesian networks in a consistent and efficient way.*

What is the basic pattern behind the problem of corresponding variables? Let us assume three random variables A,B,C, where either A is related to B or C is related to B. The conditional probabilities $P(B|A)$ and $P(B|C)$ have been estimated. The situation is much like that of a noisy-or (see Sec. 3.2.3), but the combinational function is different. The either-or decision can be modeled by a ***selection variable*** S which has two possible values $\{\tilde{a},\tilde{c}\}$ that denote the two possible corresponding variables A,C. The intended functionality can now be represented by the conditional probability $P(B|A,C,S)$ that is defined as follows:

$$P(B|A,C,S) = \left[P(b_i|a_j,c_k,s);\ i=1\ldots m, j=1\ldots n, k=1\ldots r, s\in\{\tilde{a},\tilde{c}\}\right]$$
$$\text{where } P(b_i|a_j,c_k,s) = \begin{cases} P(b_i|a_j) & \text{, if } s=\tilde{a} \\ P(b_i|c_k) & \text{, if } s=\tilde{c} \end{cases}$$

If $S=\tilde{a}$, C is irrelevant to B. If $S=\tilde{c}$, A is irrelevant to B. The resulting Bayesian network structure is presented in Fig. 3.5(a). It realizes a one-to-two mapping. A one-to-N mapping can be modeled by extending the possible values of $S\in\{\tilde{a}_1,\ldots,\tilde{a}_N\}$:

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

For the next step a new variable D is introduced that is defined analogically to B, i.e. it corresponds to either A or C (Fig. 3.5(b)). This is modeled by a new selection variable

$T \in \{\tilde{a}, \tilde{c}\}$. If the mapping of corresponding variables is exclusive this can be represented by a relation R between the two selection variables S, T:

$$\begin{aligned} &P(R|S,T) = [P(r|s,t);\ r \in \{1,0\}, s \in \{\tilde{a},\tilde{c}\}, t \in \{\tilde{a},\tilde{c}\}] \\ &\text{where } P(R=1|s,t) = \begin{cases} 1.0, & \text{if } s \neq t \\ 0.0, & \text{if } s = t \end{cases} \end{aligned}$$

The Bayesian network in Fig. 3.5(b) models an exclusive two-to-two mapping. An exclusive M-to-N mapping can be modeled by introducing M selection variables $S_1, \dots, S_M \in \{\tilde{a}_1, \dots, \tilde{a}_N\}$ with conditional probabilities:

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

and an exclusive relation R with

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

These huge probability tables need not be explicitly represented in the Bayesian network. The next section will describe an inference algorithm that combines conditioning and bucket elimination techniques in order to efficiently evaluate such networks in a general way (see 3.3.4).

Now, returning to the tennis example, the two-to-two mapping of the four evidential sets can easily be expressed (Fig. 3.6). The selection variables S, T have the values $\{\tilde{l''}_a, \tilde{l''}_b\}$. The CPTs $P(L'_{1/2}|L''_a, L''_b, S/T)$ are defined as follows:

$$\begin{aligned} &P(L'_1|L''_a, L''_b = l, S = \tilde{l''}_a) = P(L'_1|L''_a),\ l \in \Omega_L \\ &P(L'_1|L''_a = l, L''_b, S = \tilde{l''}_b) = P(L'_1|L''_b),\ l \in \Omega_L \\ &P(L'_2|L''_a, L''_b = l, T = \tilde{l''}_a) = P(L'_2|L''_a),\ l \in \Omega_L \\ &P(L'_2|L''_a = l, L''_b, T = \tilde{l''}_b) = P(L'_2|L''_b),\ l \in \Omega_L \\ &P(R=1|S,T) = [P(R=1|s,t);\ s,t \in \{\tilde{l''}_a, \tilde{l''}_b\} \\ &\text{where } \Omega_L = \{\mathit{French}, \mathit{Eastern\text{-}European}, \mathit{exotic}\}, \\ &\qquad P(R=1|s,t) = \begin{cases} 1.0, & \text{if } s \neq t \\ 0.0, & \text{if } s = t \end{cases} \end{aligned}$$

The CPTs $P(L_i|L'_i, C_i, H_i), i = 1,2$ integrate the different types of evidence. If the value of L_i differs from that of L'_i the conditional probability is set to zero:

$$P(L_i = l|L'_i = l', c, h) = \begin{cases} P(L_i = l|c,h), & \text{if } l = l' \\ 0.0, & \text{if } l \neq l' \end{cases}$$

The CPTs $P(OP'_{1/2}|OP''_a, OP''_b, S/T), P(OP_i|OP'_i, NP_i), i = 1,2$ are defined analogically.

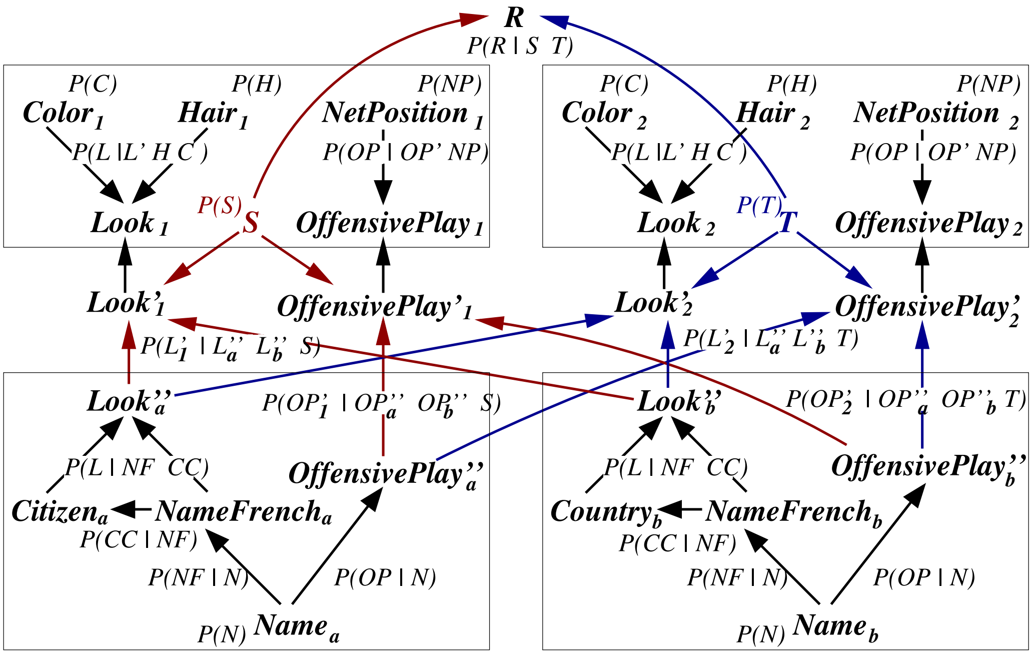


Figure 3.6: A two-to-two mapping in the tennis domain (cf. Ex. 3).

3.3 Inference in Bayesian Networks

The basic inference in a Bayesian network $\mathcal{B} = (\mathcal{V}, \mathcal{F}, \mathcal{C})$ is ***belief updating*** (bel). The values of some variables $E_1, \ldots, E_n \subset \mathcal{V}$ in the network are known and we ask about the probability of the values of another variable $Q \in \mathcal{V}$ with regard to the known evidence. The first type of variable is called observed or ***evidential variables***, and each assignment is called an observation or ***evidence***. The second type is referred to as the ***query variable***:

$$Bel(Q = q) = P(q|\mathbf{e}) = P(Q = q|E_1 = e_1, \ldots, E_n = e_n)$$

If the joint probability table is known $P(q|\mathbf{e})$ can be easily calculated by selecting the appropriate table entries and summing over all unspecified variables:

$$P(q|\mathbf{e}) = \alpha \sum_{(a_1,\ldots,a_m) \in \mathcal{V} \setminus \{Q, E_1, \ldots, E_n\}} P(a_1, \ldots, a_m, \mathbf{e}), \quad \alpha = \text{normalizing constant}$$

The second probabilistic inference that can be calculated from Bayesian networks is finding the ***most probable explanation*** (mpe). Given a set of evidences, the configuration $(a_1^*, \ldots, a_m^*)$ of all remaining variables $\{A_1, \ldots, A_n\} = \mathcal{V} \setminus \{E_1, \ldots, E_n\}$ with the maximum probability is searched for:

$$(a_1^*, \ldots, a_m^*) = \underset{(a_1,\ldots,a_m) \in \mathcal{V}/\{E_1,\ldots,E_n\}}{\operatorname{argmax}} P(a_1, \ldots, a_m, \mathbf{e})$$

The third probabilistic inference is a mixture of belief updating and finding the most probable explanation. Instead of querying the configuration of all variables of the Bayesian network, finding the ***maximum a posteriori hypothesis*** (map) requests the configuration $(b_1^*, \dots, b_k^*)$ of a subset of variables $\mathcal{Q} = \{B_1, \dots, B_k\}$. This is an mpe-task on a marginal distribution:

$$(b_1^*, \dots, b_k^*) = \underset{(b_1,\dots,b_k) \in \mathcal{Q}}{\operatorname{argmax}} \left[\sum_{(a_1,\dots,a_{m-k}) \in \mathcal{V} \setminus (\{E_1,\dots,E_n\} \cup \mathcal{Q})} P(b_1, \dots, b_k, a_1, \dots, a_{m-k}, \mathbf{e}) \right]$$

The following subsections will first concentrate on belief updating. The other tasks will be described for the bucket-elimination algorithm because their realization is very much straightforward in this framework.

Typically, the joint probability table is too large so that it is not feasible to calculate and store it. Bayesian networks represent the joint probability table in a distributed manner by using smaller conditional probability tables. Algorithms that perform inferences in this network utilize the structural characteristics of the network in order to calculate belief updates in an efficient way with short computational time and small storage requirements.

3.3.1 I-maps, moral graphs, and d-separation

All evidence does not directly influence all nodes within the network. Mostly, an evidence $E_1 = e_1$ influences a variable A only through another variable C. In such a case, we say that the variable C separates the variable E_1 from A. The consequence is that the evidences on both sides of C can be integrated independently and then be combined in the node C. The whole evidence set $\mathbf{e}$ is divided into two independent subsets $\mathbf{e} = \mathbf{e}_C' \cup \mathbf{e}_C''$ with regard to C. This concept is formalized by the ***d-separation*** property of causal networks:

Def. 1 (d-separation) *Two variables A and B in a causal network are **d-separated** if there is an intermediate variable V for all paths between A and B such that either*

- *the connection is serial or diverging and the state of V is known or*
- *the connection is converging and neither V nor any of Vs descendants have received evidence.*

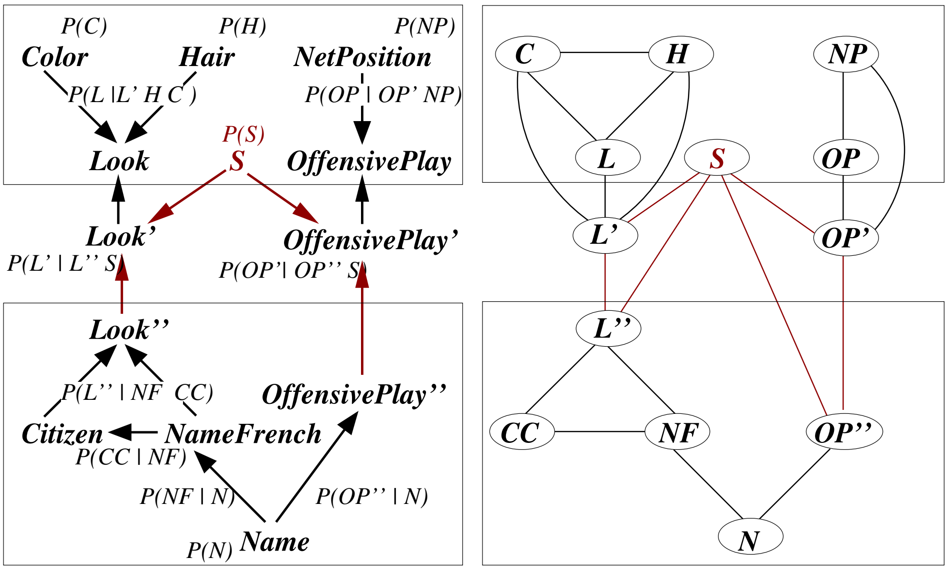
The d-separation is denoted by $\langle A|\mathcal{Z}|B\rangle$ where $\mathcal{Z}$ is the set of known variables that separate the variables A and B.

If two variables A, B are d-separated in a Bayesian network $\mathcal{B}$ with regard to $\mathcal{Z}$ the corresponding probabilities are conditionally independent:

$$\langle A|\mathcal{Z}|B\rangle_{\mathcal{B}} \quad \Rightarrow \quad P(A,B|\mathcal{Z}) = P(A|\mathcal{Z})P(B|\mathcal{Z})$$

Therefore, a Bayesian network preserves the independency assumptions of the joint probability distribution. It is a so-called ***independency map*** or ***I-map*** of P. If we can identify small subsets that d-separate the network into smaller subnets, evidences can be integrated by an efficient divide and conquer strategy.

The d-separation properties of a Baysian network can be analyzed by the construction of a so-called ***moral graph*** (see Fig. 3.7). Each node in the Bayesian network corresponds



(a) Bayesian network

(b) Moral graph

Figure 3.7: Example of a moral graph (b) that is constructed from the Bayesian network (a) of the tennis example (cf. Ex. 3, Fig. 3.6)

to a node in the moral graph. Whenever two nodes of the Bayesian network have a child-parent relation or have a common child – i.e. the random variables occur in the same conditional probability table (CPT) – they are connected by an undirected dependency edge in the moral graph.

Theorem 1 *If for all paths $\mathcal{P} = (A, C_1, \ldots, C_k, B)$ between two nodes A, B in the moral graph $\mathcal{M} = (\mathcal{V}, \mathcal{F})$ there exists a node $C \in \mathcal{Z} \subseteq \mathcal{V}$ that is an element of $\mathcal{P}$, then $\mathcal{Z}$ d-separates the nodes A, B in the corresponding Bayesian network $\mathcal{B}$: $\langle A|\mathcal{Z}|B\rangle_{\mathcal{B}}$.*

There are two properties of the moral graph that are used in the efficient organization of the belief updating task of a Bayesian network.

1. The ***width*** of a node is determined by the number of neighbors that have an edge to the node. This measure is tightly related to the complexity of the belief updating task in the Bayesian network (see 3.3.3 *bucket elimination*).

2. A subset $\mathcal{C} \subseteq \mathcal{V}$ of the nodes of the moral graph is called ***clique*** if all nodes are connected to each other. The moral graph in Fig. 3.7(b) has the cliques $\{\{L', L, C, H\}, \{L', L'', S\}, \{L'', CC, NF\}, \{NF, N\}, \{N, OP''\}, \{OP'', OP', S\}, \{NP, OP, OP'\}\}$ plus all subsets of these. The determination of the cliques of a moral graph can be used in order to organize the belief updating task in an efficient way (see 3.3.3 *junction trees*).

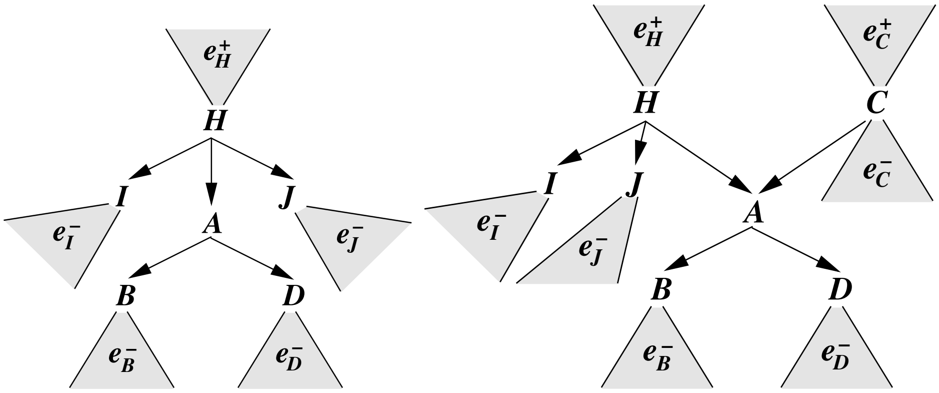


Figure 3.8: Singly connected Bayesian networks (left: tree, right: poly-tree).

3.3.2 Singly connected networks

The simplest case of belief updating is given for tree structured networks, i.e. any node in the network has only one parent. Consequently, any node A in the network d-separates its parent H and children B, D (Fig. 3.8, left). Therefore, if the belief of A shall be calculated, the evidence **e** can be divided into one ***causal subset*** (parent side) $\mathbf{e}_A^+$ and one ***diagnostic subset*** for each child $\mathbf{e}_B^-, \mathbf{e}_D^-$:

$$\mathbf{e} = \mathbf{e}_A^+ \cup \mathbf{e}_A^-, \quad \mathbf{e}_A^- = \mathbf{e}_B^- \cup \mathbf{e}_D^-$$

Pearl has proposed a recursive propagation scheme that directly exploits these independency assumptions in the network (here for the child nodes B, D):

$$Bel(a) = P(a|\mathbf{e}_A^+, \mathbf{e}_A^-) = \alpha P(\mathbf{e}_A^-|a)P(a|\mathbf{e}_A^+)$$
$$P(\mathbf{e}_A^-|a) = P(\mathbf{e}_B^-, \mathbf{e}_D^-|a) = P(\mathbf{e}_B^-|a)P(\mathbf{e}_D^-|a)$$

Conditioning over the variable B results in the conditional probability table $P(B|A)$, which is given, and the diagnostic evidential term $P(\mathbf{e}_B^-|b)$ that can be recursively computed:

$$P(\mathbf{e}_B^-|a) = \sum_b P(\mathbf{e}_B^-|b)P(b|a)$$

The causal evidence $\mathbf{e}_A^+$ can be divided into three independent subsets $\mathbf{e}_A^+ = \mathbf{e}_H^+ \cup \mathbf{e}_I^- \cup \mathbf{e}_J^-$ with regard to the parent H:

$$\begin{aligned} P(a|\mathbf{e}_A^+) &= \sum_h P(a|h)P(h|\mathbf{e}_H^+, \mathbf{e}_I^-, \mathbf{e}_J^-) \\ &= \alpha \sum_h P(a|h)P(h|\mathbf{e}_H^+)P(\mathbf{e}_I^-|h)P(\mathbf{e}_J^-|h) \end{aligned}$$

Again, the conditional probability table $P(A|H)$ is given and the other terms can be recursively computed. If the variable of an evidential term has a known value in the evidential set **e**, the term is trivially calculated by assigning 1 for the known value and 0 otherwise.

From another point of view, the recursively calculated results can be interpreted as messages from one node in the network to another node. $\pi_A(h) = P(h|e_A^+)$ is called ***causal support*** of A contributed by parent H. $\lambda_B(a) = P(\mathbf{e}_B^-|a)$ is called ***diagnostic support*** of A contributed by child B.

If the node A has more than one parent H, C in the network, but there exists no path between H and C except the one though A, the network is called a singly connected poly-tree (Fig. 3.8). In this case, a similar recursive scheme can be applied. Each diagnostic or causal subset of the parents is d-separated from the other by the complete set of parents, e.g. $\langle \mathbf{e}_H^+ | H, C | \mathbf{e}_C^+ \rangle$. Consequently, the causal support of A can be calculated by:

$$P(a|\mathbf{e}_A^+) = \alpha \sum_{h,c} P(a|h,c)P(h|\mathbf{e}_H^+)P(\mathbf{e}_I^-|h)P(\mathbf{e}_J^-|h)P(c|\mathbf{e}_C^+)P(\mathbf{e}_C^-)$$

So far, the d-separation properties could be directly inferred from the tree or poly-tree structure because there is only one path to check in any case. If there exists another path undirected loops are introduced to the network that complicate the d-separation analysis.

3.3.3 Coping with loops

If the network is not singly connected, the basic assumptions of the recursive propagation scheme are not applicable. The evidences cannot be divided in diagnostic and causal with regard to a single variable A because there might exist another path between the parents and the children of A. There are four possibilities to make propagation in networks with undirected circles tractable: (i) instead of propagating messages between nodes they are propagated between cliques of nodes that are singly connected (see ***junction trees***); (ii) additional independencies are introduced by variable assignments such that the conditioned network is singly connected (see ***conditioning***); (iii) the impact of a node on the whole network is calculated without distinguishing between diagnostic and causal support (see ***bucket elimination***); (iv) ***stochastic simulation***.

Conditioning

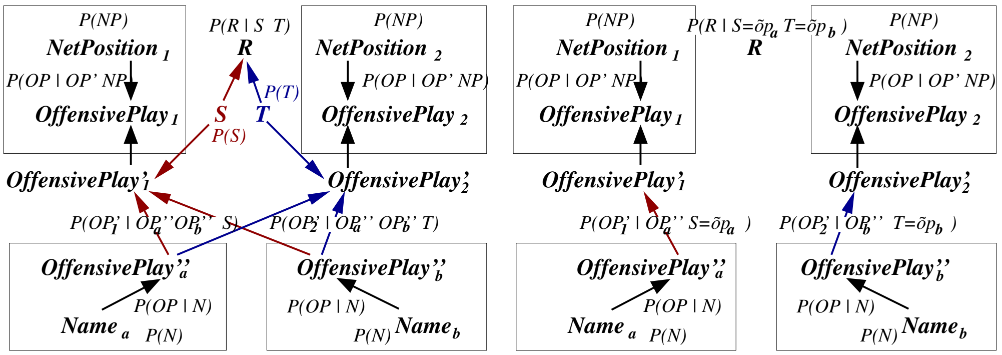
Undirected circles in a Bayesian network can be broken up by assigning a value to a variable of the circle (see Fig. 3.9). The joint probability distribution $P(\mathcal{U})$ represented by the Bayesian network is ***conditioned*** by a set of variables $C_1 \ldots C_n$:

$$P(\mathcal{U}) \xrightarrow{C_1=c^{(1)},\ldots,C_n=c^{(n)}} P(\mathcal{V}|C_1 = c^{(1)}, C_2 = c^{(n)}) \text{ with } \mathcal{V} = \mathcal{U} \setminus \{C_1, \ldots, C_n\},$$

such that $P(\mathcal{V}|C_1, \ldots, C_n)$ can be represented by a singly connected Bayesian network.

A belief updating task for a variable $A \in \mathcal{V}$ must be calculated for any configuration of the conditioning variables $C_1, \ldots, C_n$ applying e.g. the recursive propagation algorithm for singly-connected networks (cf. 3.3.2):

$$P(A|e) = \sum_{c^{(1)} \ldots c^{(n)}} P(A|e, c^{(1)} \ldots c^{(n)})\, P(c^{(1)} \ldots c^{(n)}|e)$$



(a) Bayesian network (b) Conditioning of $S,T \in \{\tilde{op}_a, \tilde{op}_b\}$

Figure 3.9: The undirected circle of the Bayesian network (a) is broken up by conditioning of the variables $S,T \in \{\tilde{op}_a, \tilde{op}_b\}$. In this example the properties of the modeling of the corresponding variables are exploited.

The $P(c^{(1)},\ldots,c^{(n)}|e)$ terms can be interpreted as a mixing weight for the configuration $(c^{(1)}\ldots c^{(n)})$. It can be easily calculated applying the rule of Bayes:

$$P(c^{(1)}\ldots c^{(n)}|e) = \alpha\, P(e|c^{(1)}\ldots c^{(n)})\, P(c^{(1)}\ldots c^{(n)}), \quad \text{where } \alpha \text{ is a normalizing factor}$$

$P(c^{(1)}\ldots c^{(n)})$ is the a priori probability of the configuration, and $P(e|c^{(1)}\ldots c^{(n)})$ can be calculated from the Bayesian network representation of $P(\mathcal{V}|C_1,\ldots,C_n)$.

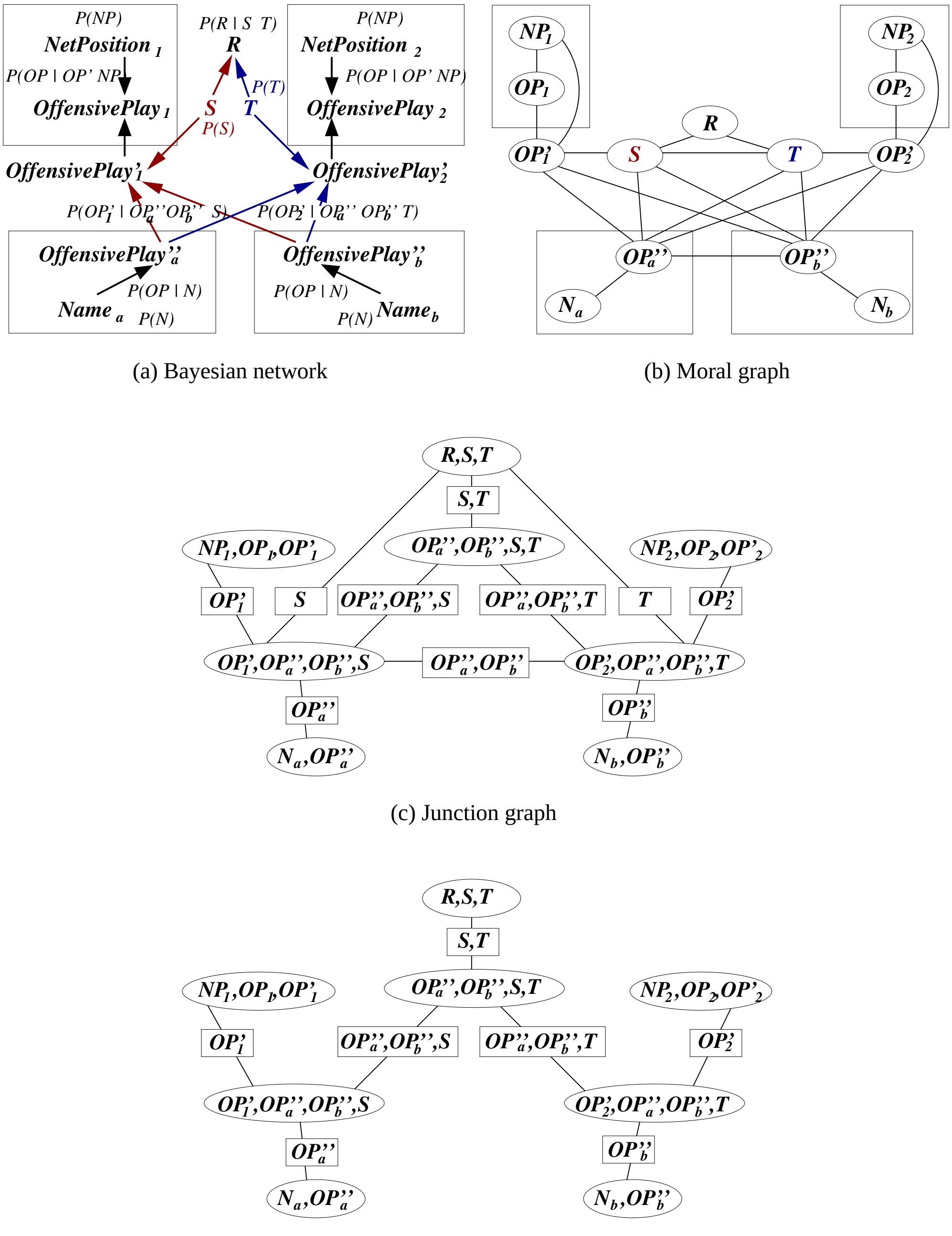
Junction trees

One way of making the d-separation explicit is the construction of a ***junction tree*** from the moral graph (see Fig. 3.10).

Def. 2 (junction graph/junction tree) *A **junction graph** for an undirected graph $\mathcal{G}$ is an undirected, labeled graph. The nodes are the cliques in $\mathcal{G}$. Every pair of nodes with a non-empty intersection has a link labeled by the intersection [Jen96, p. 85]. The **junction tree** $\mathcal{J}$ is a spanning tree of the junction graph $\mathcal{G}$, such that for each pair of nodes U,V all nodes on the path between U and V in $\mathcal{J}$ contain $U \cap V$.*

In the example shown in Fig. 3.10 eight cliques, also called ***clusters***, have been identified in the moral graph. The junction tree is constructed from the junction graph by removing appropriate edges from the graph. The set of nodes $\mathcal{Z}$ of an edge in the junction tree d-separates the sets of nodes $\mathcal{S},\mathcal{T}$ included in the two remaining subtrees: $\langle\mathcal{S}|\mathcal{Z}|\mathcal{T}\rangle$, e.g.

$$\mathcal{S} = \{NP_1, OP_1, OP'_1, OP''_a, OP''_b, N_a, S\}$$
$$\mathcal{Z} = \{OP''_a, OP''_b, S\}$$
$$\mathcal{T} = \{R, S, T, NP_2, OP_2, OP'_2, OP''_a, OP''_b, N_b\}.$$



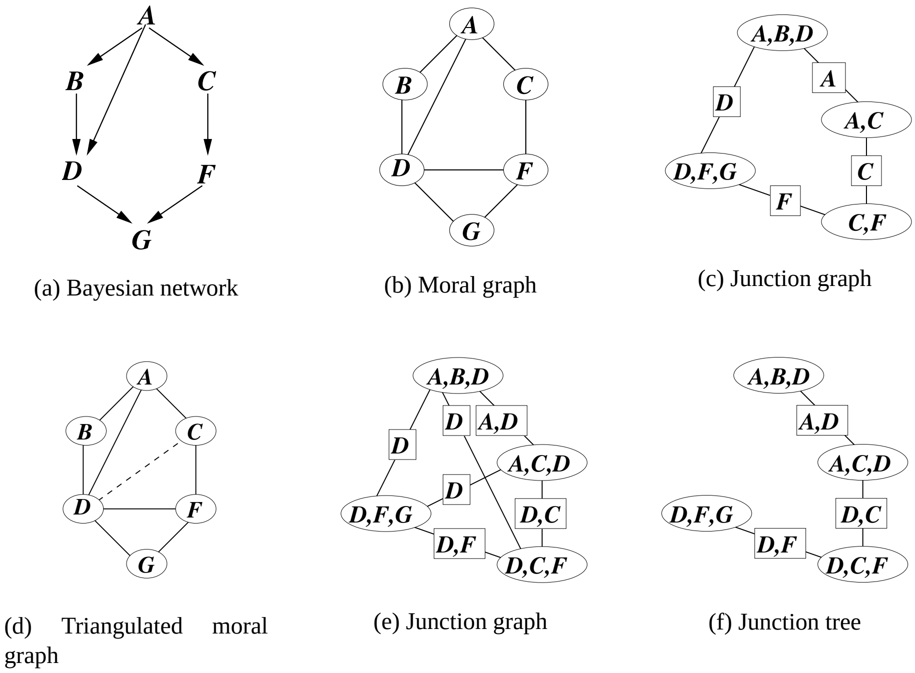
(a) Bayesian network

(b) Moral graph

(c) Junction graph

(d) Junction tree

Figure 3.10: The junction graph (c) is constructed from the cliques of the moral graph (b). The nodes are shown as ellipses. The edge labels are shown in rectangles. The junction tree defines a spanning tree of the junction graph.



(a) Bayesian network

(b) Moral graph

(c) Junction graph

(d) Triangulated moral graph

(e) Junction graph

(f) Junction tree

Figure 3.11: Triangulation: the moral graph (b) is triangulated (d). A junction tree (f) does only exist for the triangulated moral graph.

The recursive structure of the tree can now be used in order to propagate the evidence through the network.

In Fig. 3.11(a-c) the construction of the junction tree is more complicated because the junction graph has no spanning tree that complies with the properties of a junction tree. An undirected circle remains if a spanning tree is tried to build up. This problem is circumvented by the solving of a more restricted problem. An edge is added to the moral graph such that the corresponding junction graph has a junction tree. A sufficient criterion for the existence of a junction tree is that any circle of more than three nodes in the moral graph must have a ***chord***, i.e. an additional edge that connects two nodes in the circle (see Fig. 3.11(d-f)). This transformation of the moral graph is called ***triangulation***.

The junction tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ in Fig. 3.11(f) represents the joint probability distribution $P(A,B,C,D,F,G)$. The nodes $\mathcal{V} = \{W_1, \ldots, W_n\}$ are labeled with ***cluster tables*** $\mathbf{t}_{W_i}$, the edges $\mathcal{E} = \{S_1, \ldots, S_m\}$ with ***separator tables*** $\mathbf{t}_{S_j}$. The joint probability distribution $P(U)$ can be calculated as the product of all cluster tables divided by the product of all separator tables:

$$P(U) = P(A,B,C,D,F,G) = \frac{\prod_{i=1\ldots n} \mathbf{t}_{W_i}}{\prod_{j=1\ldots m} \mathbf{t}_{S_j}} \tag{3.2}$$

Initially each node cluster W_i and edge separator S_j is assigned a table of ones, e.g. $\mathbf{t}_{W_i} = (1,1,\ldots,1)$. Then each CPT $P(X|Y_1,\ldots,Y_k)$ is multiplied by the table $\mathbf{t}_{W_i}$ of an unique cluster W_i with $\{X,Y_1,\ldots,Y_k\} \subseteq W_i$. Evidence $\mathbf{e} = \{A = a_2, D = d_1\}$ is represented by the corresponding tables $\underline{e}_A = (0,1,0,\ldots,0), \underline{e}_D = (1,0,\ldots,0)$. These are analogically inserted into the junction tree resulting in a representation of

$$P(U,\mathbf{e}) = P(A = a_2, B, C, D = d_1, F, G) = \frac{\prod_{i=1\ldots n} \mathbf{t}_{W_i}}{\prod_{j=1\ldots m} \mathbf{t}_{S_j}}. \tag{3.3}$$

The propagation in junction trees is based on an operation called ***absorption*** that rearranges the information represented in the junction tree so that it remains invariant under Eq. 3.3.

Def. 3 (*absorption*) *Let V and W be neighbors in a junction tree, let S be their separator, and let $\mathbf{t}_V, \mathbf{t}_W$ and $\mathbf{t}_S$ be their tables. The* ***absorption*** *operation is the result of the following procedure [Jen96, p. 73]:*

- *calculate $\mathbf{t}_S^* = \sum_{V \setminus S} \mathbf{t}_V$;*
- *give S the table $\mathbf{t}_S^*$;*
- *give W the table $\mathbf{t}_W^* = \mathbf{t}_W \frac{\mathbf{t}_S^*}{\mathbf{t}_S}$.*

We say that W has *absorbed* from V. After the propagation is complete, i.e. the tables $\mathbf{t}_{W_i}, \mathbf{t}_{S_j}$ remain invariant under absorption, we have for each node V in the junction tree and each separator S:

$$\mathbf{t}_V = \sum_{U \setminus V} P(U,\mathbf{e}) = P(V,\mathbf{e}) \text{ and } \mathbf{t}_S = P(S,\mathbf{e}) \tag{3.4}$$

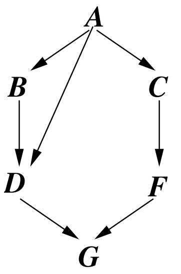
From this representation the belief $P(X|\mathbf{e})$ of a node $X \in V$ can easily be calculated:

$$P(X|\mathbf{e}) = \alpha \sum_{V \setminus X} \mathbf{t}_V, \quad \text{where } \alpha \text{ is a normalizing factor} \tag{3.5}$$

The propagation algorithm can be realized by a recursive scheme that is analogical to the propagation in singly-connected networks. The complexity of the algorithm is determined by the sizes of the clusters and separation tables that are exponential in the number of nodes of the cliques. In the worst case, the junction tree consists of a single clique containing all nodes of the Bayesian network. The number of absorption needed is linear to the number of cliques.

Bucket elimination

Bucket elimination is a general problem solving method that is tightly related to the dynamic programming approach. The application to probabilistic inference was introduced by Rina Dechter [Dec98]. The bucket elimination scheme shares some ideas with junction trees but is organized in a simpler linear way instead of operating on graphs.



(a) Bayesian network

$\mathcal{G}: P(G\|D,F)$	$G = g_2$	
$\mathcal{F}: P(F\|C)$		$\lambda_G(D,F)$
$\mathcal{B}: P(B\|A)\ P(D\|B,A)$	$B = b_1$	
$\mathcal{D}:$		$\lambda_F(D,C)\ \lambda_B(D,A)$
$\mathcal{A}: P(C\|A)\ P(A)$		$\lambda_D(A,C)$
$\mathcal{C}:$		$\lambda_A(C)$

(b) Buckets

Figure 3.13: Bucket scheme for the Bayesian network in (a) and the variable ordering C, A, D, B, F, G. (b) shows the buckets after they have been processed in reverse order.

Assume the network presented in Fig. 3.13(a). The belief updating task for variable C is defined as follows if evidence $\mathbf{e} = \{G = g_2, B = b_1\}$ has been observed:

$$
\begin{aligned}
P(c|G = g_2, B = b_1) &= \alpha \sum_{f,d,a} P(c, f, d, g_2, a, b_1) \\
&= \alpha \sum_{f,d,a} P(g_2|d, f)\ P(d|b_1, a)\ P(b_1|a)\ P(f|c)\ P(c|a)
\end{aligned}
$$

The same term can be written in the algebra of conditional probability tables with $\underline{e}_G = (0, 1, 0, \ldots, 0), \underline{e}_B = (1, 0, \ldots, 0)$:

$$
\begin{aligned}
&P(C|G = g_2, B = b_1) \\
&\quad = \alpha \sum_{a,d,b,f,g} P(G|D,F)\ P(D|B,A)\ P(B|A)\ P(F|C)\ P(C|A)\ P(A)\ \underline{e}_G\ \underline{e}_B \\
&\quad = \alpha \sum_a P(A) \sum_d \sum_b P(D|B,A)\ P(B|A)\ \underline{e}_B \sum_f P(F|C)\ \sum_g P(G|D,F)\ \underline{e}_G
\end{aligned}
$$

The splitting of the complete summation into summations over single variables is essentially what the bucket scheme introduces. Conditional probability tables are factored out. The scope of the summations then defines the ***buckets***. The bucket operation consists of a multiplication of all CPTs in scope and a summation over the ***bucket variable***. If the bucket variable has been observed the summation can be substituted by a selection of the subtable that is determined by the observed value. The expression is evaluated starting with the innermost bucket. The result of a bucket $\lambda_X(Y_1, \ldots, Y_N)$ is a new CPT that may be factored out again:

$$
\begin{aligned}
&\alpha \sum_a P(A) \sum_d \sum_b P(D|B,A)\; P(B|A)\; \underline{e}_B \sum_f P(F|C)\; \sum_g P(G|D,F)\; \underline{e}_G \\
=&\alpha \sum_a P(A) \sum_d \sum_b P(D|B,A)\; P(B|A)\; \underline{e}_B \sum_f P(F|C)\; \lambda_G(D,F) \\
=&\alpha \sum_a P(A) \sum_d \lambda_F(D,C) \sum_b P(D|B,A)\; P(B|A)\; \underline{e}_B \\
=&\alpha \sum_a P(A) \sum_d \lambda_F(D,C)\; \lambda_B(D,A) \\
=&\alpha \sum_a P(A)\; \lambda_D(C,A) \\
=&\alpha\; \lambda_A(C)
\end{aligned}
$$

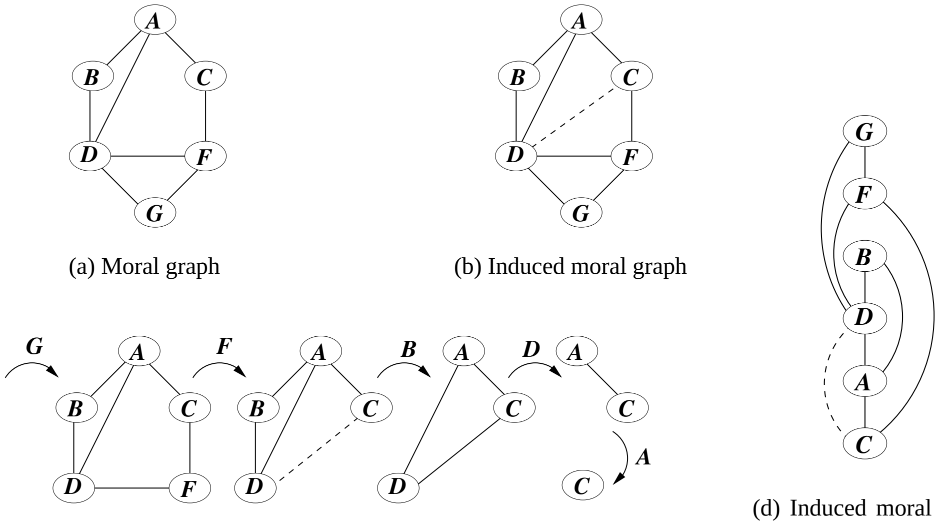
In each step, one bucket is *eliminated* from the expression. Fig. 3.13(b) explicitly shows the representation of the buckets. Initially, the CPTs $P(X|Y_1,\ldots,Y_k)$ of the Bayesian network in Fig. 3.13(a) are inserted into the buckets. A CPT is placed in the first bucket whose bucket variable is member of the CPT. The buckets are checked in reverse order. For example, $P(B|A)$ is placed in bucket $\mathcal{B}$. Then the evidences, e.g. $G = g_2$, are added to the corresponding bucket, here $\mathcal{G}$. Afterwards, the buckets are evaluated in reverse order. For example, bucket $\mathcal{B}$ produces a message $\lambda_B(D,A)$ that is placed in bucket $\mathcal{D}$. The result of the belief updating task is collected in the last bucket, that is the bucket of the query variable. It is calculated by a normalized product over all elements in the bucket.

The complexity of the algorithm strongly depends on the ordering of the buckets. A bad ordering would for example start with bucket D:

$$\lambda_D(G,F,A,B) = \sum_d P(G|D,F)\; P(D|B,A)$$

resulting in a large CPT with a dimension that is the product of the variable dimensions, i.e. the number of possible values of each variable.

The calculation of the optimal variable ordering, with the smallest CPTs, is NP-hard. However, a good solution can be obtained by analysis of the moral graph using a greedy strategy (see Fig. 3.14). The moral graph of the Bayesian network in Fig. 3.12(a) is shown in Fig. 3.14(a). The ordering of variables is obtained by selecting the node with minimal width in the moral graph, here one of the nodes B,C,G. In Fig. 3.14(c) G is selected and eliminated from the moral graph. An elimination step consists of the deletion of the node and the insertion of edges between all nodes that were connected to G. In the next step, node F is selected and eliminated resulting in the insertion of an edge between D and C. The algorithm continues with the variables B,D,A until only the query variable C is left, thereby calculating the ordering C,A,D,B,F,G. The moral graph with all inserted edges during selection defines the ***induced moral graph*** (Fig. 3.14(b)). It equals the triangulated moral graph used for junction trees (see Fig. 3.11(d)). The ***induced width*** of a node in the moral graph with regard to the ordering C,A,D,B,F,G is defined as the number of edges to earlier neighbors in the induced moral graph (Fig. 3.14(d)). Earlier means *before* with regard to the ordering. The induced width w of a node X is identical to the size of the



(a) Moral graph

(b) Induced moral graph

(c) Elimination

(d) Induced moral graph with regard to the ordering C, A, D, B, F, G.

Figure 3.14: Finding a good variable ordering based on the minimal induced width.

message $\lambda_X(Y_1,\ldots,Y_w)$ that is send from the bucket X to an earlier one. Consequently, it is a measure for the complexity of the belief updating task.

The finding of the ***most probable explanation*** (mpe) and the ***maximum a posteriori hypothesis*** (map) can be easily formulated in the bucket elimination framework. In the first case all bucket operations are changed from summation to maximization. In the second case only those variables that belong to the hypothesis are maximized:

- most probable explanation:

$$\begin{aligned}
&\max_{c,a,d,f} P(C,A,D,F,G=g_2,B=b_1)\\
&= \max_{c,a,d,b,f,g} P(G|D,F)\; P(D|B,A)\; P(B|A)\; P(F|C)\; P(C|A)\; P(A)\; \underline{e}_G\; \underline{e}_B\\
&= \max_{c}\; \max_{a} P(A) \max_{d}\; \max_{b} P(D|B,A)\; P(B|A)\; \underline{e}_B \max_{f} P(F|C)\; \max_{g} P(G|D,F)\; \underline{e}_G
\end{aligned}$$

- maximum a posteriori hypothesis for the variables C,F:

$$\begin{aligned}
&max_{c,f} P(C,F|G=g_2,B=b_1)\\
&= \max_{c,f} \alpha \sum_{a,d,b,g} P(G|D,F)\; P(D|B,A)\; P(B|A)\; P(F|C)\; P(C|A)\; P(A)\; \underline{e}_G\; \underline{e}_B\\
&= \max_{c}\; \max_{f} P(F|C)\; \alpha \sum_{a} P(A) \sum_{d} \sum_{b} P(D|B,A)\; P(B|A)\; \underline{e}_B \sum_{g} P(G|D,F)\; \underline{e}_G
\end{aligned}$$

Each maximization stores the selected argmax value for each resulting table entry. The most probable explanation or maximum a posteriori hypothesis is collected during back tracking the last selected maximum value.

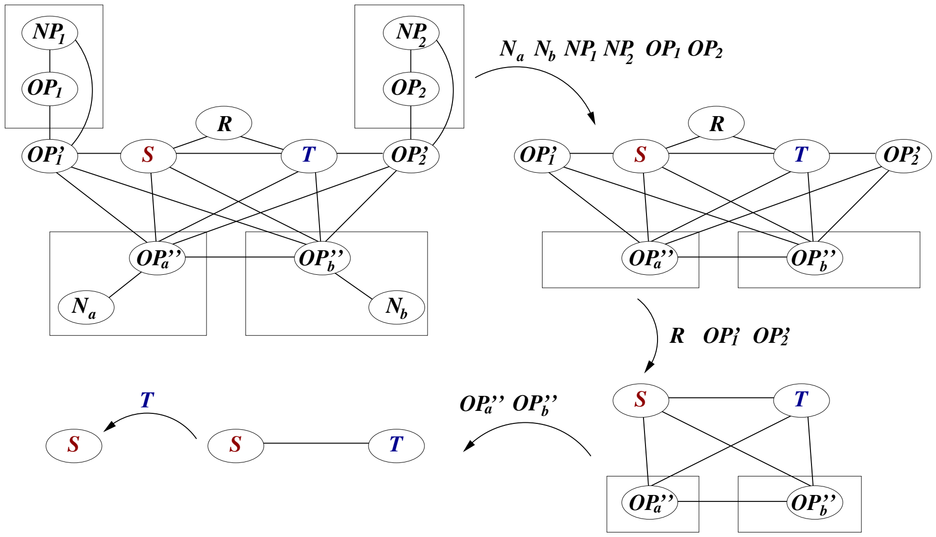


Figure 3.15: Calculating the variable ordering for the tennis example (see Fig. 3.10(a)).

The variable ordering for the map-task must be changed slightly because the summations have to be processed before the maximum-operations. Therefore, the algorithm for generating the ordering is first applied to the summation variables (selecting A,D,B,G) and then applied to the maximization variables (selecting F,C).

In the tennis example with Ronald Agenor and Ivan Lendl (Ex. 3) the correct mapping between the two players on the court and the two names on the scoreboard must be found. This problem can be formulated as a map-task, that is finding the maximum a posteriori hypothesis of the two selection variables S,T (cf. Fig. 3.6). In the following the simplified tennis network from Fig. 3.10(a) will be discussed. Its moral graph is given in Fig. 3.15. The greedy algorithm for calculating the variable ordering first selects the variables N_a,N_b that have the induced width $w=1$, and continues with $NP_1,NP_2(w=2)$ and $OP_1,OP_2(w=1)$. Then R is selected with $w=2$ and $OP'_1,OP'_2(w=3)$. The last summation variables OP''_a,OP''_b have the induced width $w=3$ and $w=2$, respectively. As the first maximization variable, T is eliminated with $w=1$ leaving the last maximization variable S. Thus, the resulting ordering is $S,T,OP''_a,OP''_b,OP'_1,OP'_2,R,OP_1,OP_2,NP_1,NP_2,N_a,N_b$.

The bucket elimination scheme for this ordering is shown in Fig. 3.16. The complexity strongly depends on the message calculation of the buckets $\mathcal{OP}'_2,\mathcal{OP}'_1,\mathcal{OP}''_b,\mathcal{OP}''_a$, i.e. those of the corresponding variables. For an N-to-M mapping the message size is exponential in N and M. If variables $X_1,\ldots,X_N$ shall be mapped to variables $Y_1,\ldots,Y_M$, N selection variables with M possible values are needed: $S_1,\ldots,S_N \in \{\tilde{y}_1,\ldots,\tilde{y}_M\}$. Buckets $\mathcal{X}_i$ calculate the messages $\lambda_{X_i}(Y_1,\ldots,Y_M,S_i)$ that are collected in Bucket $\mathcal{Y}_M$. Bucket $\mathcal{Y}_M$ then generates $\lambda_{Y_M}(Y_1,\ldots,Y_{M-1},S_1,\ldots,S_N)$ whose size is exponential in N and M. For larger Ns and Ms this is not tractable. Section 3.3.4 will present a solution that combines bucket elimination and conditioning techniques.

$$
\begin{array}{rll}
\mathcal{N}_b: & P(OP''_b|N_b)\;P(N_b) & N_b = \mathit{Lendl} \\
\mathcal{N}_a: & P(OP''_a|N_a)\;P(N_a) & N_a = \mathit{Agenor} \\
\mathcal{NP}_2: & P(OP_2|OP'_2,NP_2)\;P(NP_2) & NP_2 = \mathit{false} \\
\mathcal{NP}_1: & P(OP_1|OP'_1,NP_1)\;P(NP_1) & NP_1 = \mathit{true} \\
\mathcal{OP}_2: & & \lambda_{NP_2}(OP_2,OP'_2) \\
\mathcal{OP}_1: & & \lambda_{NP_1}(OP_1,OP'_1) \\
\mathcal{R}: & P(R|S,T) & R = 1 \\
\mathcal{OP}'_2: & P(OP'_2|OP''_a,OP''_b,T) & \lambda_{NP_2}(OP'_2) \\
\mathcal{OP}'_1: & P(OP'_1|OP''_a,OP''_b,S) & \lambda_{NP_1}(OP'_1) \\
\mathcal{OP}'_b: & & \lambda_{N_b}(OP''_b)\;\lambda_{OP'_2}(OP''_a,OP''_b,T)\;\lambda_{OP'_1}(OP''_a,OP''_b,S) \\
\mathcal{OP}'_a: & & \lambda_{N_a}(OP''_a)\;\lambda_{OP'_b}(OP'_a,S,T) \\
\mathcal{T}: & P(T) & \lambda_R(S,T)\lambda_{OP'_a}(S,T) \\
\mathcal{S}: & P(S) & \lambda_T(S)
\end{array}
$$

Figure 3.16: Bucket elimination for the tennis example (cf. Fig. 3.10(a)) with regard to the variable ordering $S,T,OP''_a,OP''_b,OP'_1,OP'_2,R,OP_1,OP_2,NP_1,NP_2,N_a,N_b$.

Stochastic Simulation

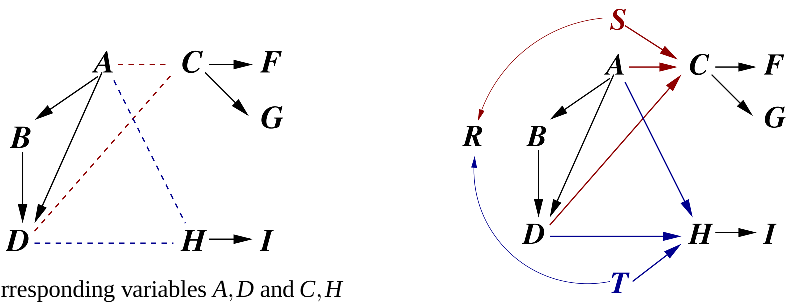
The approach of stochastic simulation is completely different to the analytical methods discussed so far. They do not provide exact inferences. Instead, the result of a belief updating task is approximated by performing simulation runs. The Bayesian network is used to generate random samples, i.e. possible configurations of the modeled variables. The probability of any event or combination of events can then be computed by counting the percentage of samples in which the event is true [Pea88, pp. 210].

The topic of approximate inference in Bayesian networks will not be discussed any further because it turns out that exact inference is possible with regard to the scope of this thesis. The interested reader may refer to the relevant literature, e.g. [Pea88, WKt$^+$99, Dec97].

3.3.4 A conditional bucket elimination scheme

As demonstrated in section 3.3.3, the evaluation of the Bayesian network in the tennis example (3.9) can be considerably simplified if conditioning is performed over the selection variables S,T. In general it is a difficult problem to decide which variables are good candidates for applying the conditioning technique. In this case, it is just straightforward because of the definition of the conditional probability tables that model the mapping of corresponding variables (cf. Sec. 3.2.4):

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



(a) Corresponding variables A,D and C,H

(b) Bayesian network

Figure 3.17: Bayesian network with corresponding variables A,D and C,H.

On the other hand, bucket elimination is a very simple and general technique for evaluating arbitrarily structured Bayesian networks. The idea to combine both techniques for an efficient solution of the correspondence problem in multi-modal processing is obvious.

Beside this contribution, other researchers proposed several approaches for combining bucket elimination and conditioning [DR96, Dec96, EFD96]. In the following a novel approach will be presented that employs the conditioning technique only on those parts of the network that benefit from the mapping properties.

Conditional buckets

In order to capture the conditioning technique in an extended bucket elimination scheme, the concept of ***conditional buckets*** will be introduced. Let $P(A,B,C,D,F,G,H,I); A,B,C,D,F,G,H,I \in \{0,1\}$ be the joint probability distribution of the modeled domain. There are two variables A,D that correspond to one of the variables C,H (Fig. 3.17(a)), i.e. an exclusive two-to-two mapping. This can be modeled by two selection variables $S,T \in \{\tilde{a},\tilde{d}\}$ (Fig. 3.17(b)). The finding of the maximum a posteriori hypothesis (s^*,t^*) given evidence $\mathbf{e} = \{B=1, F=0, I=1\}$ can be formulated as follows:

$$\begin{aligned}
(s^*,t^*) &= \operatorname*{argmax}_{s,t} \left[\sum_{a,c,d,g,h} P(s,t,a,c,d,g,h|B=1,F=0,I=1) \right] \\
&= \operatorname*{argmax}_{s,t} P(R=1|s,t)\, P(s)\, P(t) \cdot \\
&\quad \alpha \sum_{a,c,d,g,h} P(F=0|c)\, P(g|c)\, P(c|a,d,s)\, P(I=1|h)\, P(h|a,d,t) \cdot \\
&\qquad \cdot P(d|B=1,a)\, P(B=1|a)\, P(a)
\end{aligned}$$

The general formula of the conditioning technique (cf. Sec. 3.3.3) is

$$P(A|e) = \sum_{c^{(1)}\ldots c^{(n)}} P(A|e, c^{(1)} \ldots c^{(n)})\, P(c^{(1)} \ldots c^{(n)}|e)$$

Here, the conditioning variables themselves are the query variables. Therefore, only

$$P(c^{(1)}\dots c^{(n)}|e) = \alpha\, P(e|c^{(1)}\dots c^{(n)})\, P(c^{(1)}\dots c^{(n)}), \quad \text{where } \alpha \text{ is a normalizing factor}$$

has to be calculated. The task has been simplified because $P(e|c^{(1)}\dots c^{(n)})$ can be calculated on a simplified Bayesian network. Applying this on the Bayesian network in Fig. 3.17(b) reveals that different net structures are obtained from different values of the conditioning variables:

$$\begin{aligned}
&P(B=1,F=0,I=1|S=s_j,T=t_k)\\
&= \sum_{a,c,d,g,h} P(F=0|c)\, P(g|c) \begin{Bmatrix} P(c|a,S=s_j), & \text{if } s_j=\tilde{a} \\ P(c|d,S=s_j), & \text{if } s_j=\tilde{d} \end{Bmatrix} \cdot \\
&\qquad \cdot P(I=1|h) \begin{Bmatrix} P(h|a,T=t_k), & \text{if } t_k=\tilde{a} \\ P(h|d,T=t_k), & \text{if } t_k=\tilde{d} \end{Bmatrix} P(d|B=1,a)\, P(B=1|a)\, P(a)
\end{aligned} \tag{3.6}$$

Therefore, parts of the Bayesian network can only be evaluated once. Other parts of the network must be evaluated separately. This can be managed by introducing the concept of ***conditional buckets***.

Def. 4 (conditional bucket) *A **conditional bucket** $\left[\mathcal{Y}|X=x\right]$ receives all those CPTs $P(A_1,\dots,A_n|Y,B_1,\dots,B_m,X = x)$, $P(Y,A_1,\dots,A_n|B_1,\dots,B_m,X = x)$, or $\lambda_C(Y,A_1,\dots,A_n,X=x)$ that contain the variable Y and are conditioned by $X=x$.*
*The **conditioned message** of the bucket $\left[\mathcal{Y}|X=x\right]$ is defined with regard to the bucket operation $func \in \{\sum,\max\}$ as:*

$$\begin{aligned}
\lambda_Y(A_1,\dots,A_n,X=x) &= \underset{y}{func}\,(\prod_j CPT_j^{(Y,X=x)})\,\lambda_Y(Y,B_1,\dots,B_m)\\
\textit{where}\quad \lambda_Y(Y,B_1,\dots,B_m) &= \prod_k CPT_k^{(Y)},\\
CPT_j^{(Y,X=x)} &\in \left[\mathcal{Y}|X=x\right],\\
CPT_k^{(Y)} &\in \mathcal{Y}, \quad \textit{where } \mathcal{Y} \textit{ is the unconditioned bucket.}
\end{aligned}$$

The probabilities that shall be calculated are given by Eq. 3.6. Therefore, the probability tables $P(F|C), P(G|C), P(C|A,S=\tilde{a}), P(C|D,S=\tilde{d}), P(I|H), P(H|A,T=\tilde{a}), P(H|D,T=\tilde{d}), P(D|B,A), P(B|A), P(A)$ and evidential tables $\underline{e}_B = (1.0,\ 0.0), \underline{e}_F = (0.0,\ 1.0)$, and $\underline{e}_I = (1.0,\ 0.0)$ have to be inserted into the bucket scheme. Given the variable ordering

D,A,H,C,B,I,G,F, the resulting bucket allocation is:

$$\begin{aligned}
&\mathcal{F} : P(F|C)\ \underline{e}_F\\
&\mathcal{G} : P(G|C)\\
&I : P(I|H)\ \underline{e}_I\\
&\mathcal{B} : P(D|B,A)\ P(B|A)\ \underline{e}_B\\
&\mathcal{C} : \begin{cases} [\mathcal{C}|S=\tilde{a}] : & P(C|A,S=\tilde{a})\\ [\mathcal{C}|S=\tilde{d}] : & P(C|D,S=\tilde{d}) \end{cases}\\
&\mathcal{H} : \begin{cases} [\mathcal{H}|T=\tilde{a}] : & P(H|A,T=\tilde{a})\\ [\mathcal{H}|T=\tilde{d}] : & P(H|D,T=\tilde{d}) \end{cases}\\
&\mathcal{A} : P(A)\\
&\mathcal{D} :
\end{aligned}$$

The buckets $\mathcal{F}$ to $\mathcal{B}$ can be processed in the normal way. Elimination of these buckets results in:

$$\begin{aligned}
&\mathcal{C} : \lambda_F(C)\ \lambda_G(C) \begin{cases} [\mathcal{C}|S=\tilde{a}] : & P(C|A,S=\tilde{a})\\ [\mathcal{C}|S=\tilde{d}] : & P(C|D,S=\tilde{d}) \end{cases}\\
&\mathcal{H} : \lambda_I(H) \begin{cases} [\mathcal{H}|T=\tilde{a}] : & P(H|A,T=\tilde{a})\\ [\mathcal{H}|T=\tilde{d}] : & P(H|D,T=\tilde{d}) \end{cases}\\
&\mathcal{A} : P(A)\ \lambda_B(D,A)\\
&\mathcal{D} :
\end{aligned}$$

Now, the messages of the unconditioned and conditioned parts of $\mathcal{C}$ are calculated:

$$\begin{aligned}
&\lambda_C(A,S=\tilde{a}) = \sum_c P(C|A,S=\tilde{a})\ \lambda_C(C)\\
&\lambda_C(D,S=\tilde{d}) = \sum_c P(C|D,S=\tilde{d})\ \lambda_C(C)\\
&\text{where } \lambda_C(C) = \lambda_F(C)\ \lambda_G(C).
\end{aligned}$$

and re-inserted:

$$\begin{aligned}
&\mathcal{H} : \lambda_I(H) \begin{cases} [\mathcal{H}|T=\tilde{a}] : & P(H|A,T=\tilde{a})\\ [\mathcal{H}|T=\tilde{d}] : & P(H|D,T=\tilde{d}) \end{cases}\\
&\mathcal{A} : P(A)\ \lambda_B(D,A) \begin{cases} [\mathcal{A}|S=\tilde{a}] : & \lambda_C(A,S=\tilde{a})\\ [\mathcal{A}|S\neq\tilde{a}] : & \end{cases}\\
&\mathcal{D} : \begin{cases} [\mathcal{D}|S=\tilde{d}] : & \lambda_C(D,S=\tilde{d})\\ [\mathcal{D}|S\neq\tilde{d}] : & \end{cases}
\end{aligned}$$

Note that in the buckets $\mathcal{A}$ and $\mathcal{D}$ the insertion of the messages $\lambda_C(A,S=\tilde{a})$ and $\lambda_C(D,S=\tilde{d})$, respectively, results in the introduction of *two* conditional buckets that cover all possible assignments of the conditioning variable. Then the $\mathcal{H}$ buckets are processed:

$$\mathcal{A}: P(A)\,\lambda_B(D,A)\begin{cases}[\mathcal{A}|S=\tilde{a}]: & \lambda_C(A,S=\tilde{a})\begin{cases}[\mathcal{A}|T=\tilde{a}]: & \lambda_H(A,T=\tilde{a})\\ [\mathcal{A}|T\neq\tilde{a}]: & \end{cases}\\ [\mathcal{A}|S\neq\tilde{a}]: & \begin{cases}[\mathcal{A}|T=\tilde{a}]: & \lambda_H(A,T=\tilde{a})\\ [\mathcal{A}|T\neq\tilde{a}]: & \end{cases}\end{cases}$$

$$\mathcal{D}:\begin{cases}[\mathcal{D}|S=\tilde{d}]: & \lambda_C(D,S=\tilde{d})\begin{cases}[\mathcal{D}|T=\tilde{d}]: & \lambda_H(D,T=\tilde{d})\\ [\mathcal{D}|T\neq\tilde{d}]: & \end{cases}\\ [\mathcal{D}|S\neq\tilde{d}]: & \begin{cases}[\mathcal{D}|T=\tilde{d}]: & \lambda_H(D,T=\tilde{d})\\ [\mathcal{D}|T\neq\tilde{d}]: & \end{cases}\end{cases}$$

The buckets $\mathcal{A},\mathcal{D}$ are conditioned over both variables S,T with two cases each so that four messages have to be calculated. Due to the tree-structured representation that has been chosen here for simplicity reasons, the message $\lambda_H(A,T=\tilde{a})$ has to be duplicated. A message $\lambda_A(D,S=\tilde{a},T\neq\tilde{a})$ has to be calculated as follows:

$$\begin{aligned}\lambda_A(D,S=\tilde{a},T\neq\tilde{a}) &= \underset{a}{func}\left(\prod_i CPT_i^{(A,S=\tilde{a},T\neq\tilde{a})}\right)\lambda_A(A,S=\tilde{a})\,\lambda_A(A,T\neq\tilde{a})\,\lambda_A(A,D)\\ &= \sum_a \lambda_C(A,S=\tilde{a})\,P(A)\,\lambda_B(D,A)\\ \text{where } \lambda_A(A,S=\tilde{a}) &= \prod_j CPT_j^{(A,S=\tilde{a})} = \lambda_C(A,S=\tilde{a}),\\ \lambda_A(A,T\neq\tilde{a}) &= \prod_k CPT_k^{(A,T\neq\tilde{a})} = (1,\ldots,1),\\ \lambda_A(A,D) &= \prod_l CPT_l^{(A)} = P(A)\,\lambda_B(D,A).\end{aligned}$$

The evaluation of $\mathcal{A}$ results in:

$$\mathcal{D}:\begin{cases}[\mathcal{D}|S=\tilde{d}]: & \lambda_C(D,S=\tilde{d})\begin{cases}[\mathcal{D}|T=\tilde{a}]: & \lambda_A(D,S\neq\tilde{a},T=\tilde{a})\\ [\mathcal{D}|T=\tilde{d}]: & \lambda_H(D,T=\tilde{d})\,\lambda_A(D,S\neq\tilde{a},T\neq\tilde{a})\end{cases}\\ [\mathcal{D}|S=\tilde{a}]: & \lambda_A(D,S=\tilde{a})\begin{cases}[\mathcal{D}|T=\tilde{a}]: & \lambda_A(D,S=\tilde{a},T=\tilde{a})\\ [\mathcal{D}|T=\tilde{d}]: & \lambda_H(D,T=\tilde{d})\,\lambda_A(D,S=\tilde{a},T\neq\tilde{a})\end{cases}\end{cases}$$

Finally, $\mathcal{D}$ is eliminated. The results are collected in the $\mathcal{S}$- and $\mathcal{T}$-buckets:

$$\begin{aligned}&\mathcal{R}: P(R|S,T)\,\underline{e}_R\\ &\mathcal{T}: P(T)\begin{bmatrix}\lambda_D(S=\tilde{a},T=\tilde{a}) & \lambda_D(S=\tilde{a},T=\tilde{d})\\ \lambda_D(S=\tilde{d},T=\tilde{a}) & \lambda_D(S=\tilde{d},T=\tilde{d})\end{bmatrix}\\ &\mathcal{S}: P(S)\end{aligned}$$

The remaining buckets can be processed using the normal bucket-scheme.

So far, the exclusive criterion that is modeled in the CPT $P(R|S,T)$ is not exploited during elimination of the buckets. A closer look at it reveals that $P(R=1|S=s,T=t) = 0.0$, if $s=t$ and thus

$$P(R=1,s,t,\mathbf{e}) = \alpha\, P(R=1|s,t)\; P(s)\; P(t)\; P(\mathbf{e}|s,t) = 0.0, \text{if } s=t.$$

Therefore, $P(\mathbf{e}|s,t)$ need not be calculated in this case. This idea can be integrated into the conditional bucket scheme by discarding any conditional buckets $[\mathcal{X}|S=s_i,T=s_i]$ whose conditioning variables S,T have been assigned the same value $s_i \in \{\tilde{a},\tilde{d}\}$. The messages $\lambda_X(S=s_i,T=s_j)$ are then collected in a combined bucket $\mathcal{S},\mathcal{T}$ that is processed by an exclusive maximum operation over the variables S,T.

Def. 5 (exclusive bucket) *An **exclusive bucket** $[\mathcal{S}_1,\ldots,\mathcal{S}_n]$ receives all CPTs that contain one of the variables $S_1,\ldots,S_n \in \{s_1,\ldots,s_m\}$. The elimination of the bucket is performed by an exclusive maximum operation:*

$$\lambda_{S_1,\ldots,S_n}(X_1,\ldots,X_m) = \max_{(s^{(1)},\ldots,s^{(n)}),s^{(i)}\neq s^{(j)},i\neq j,\, i,j\in\{1\ldots n\}} \left(\prod_{k_1} CPT_{k_1}^{(S_1)}\right)\cdots\left(\prod_{k_n} CPT_{k_n}^{(S_n)}\right)$$

Substituting the buckets $\mathcal{R},\mathcal{S},\mathcal{T}$ by the exclusive bucket $[\mathcal{S},\mathcal{T}]$ yields:

$$[\mathcal{S},\mathcal{T}] : P(S)\; P(T) \begin{bmatrix} 0 & \lambda_D(S=\tilde{a},T=\tilde{d}) \\ \lambda_D(S=\tilde{d},T=\tilde{a}) & 0 \end{bmatrix}$$

The evaluation is performed by:

$$(s^*,t^*) = \operatorname*{argmax}_{s,t,\, s\neq t} P(S)\; P(T)\; \lambda_D(S=s,T=t)$$

Although, the sizes of the propagated messages have been substantially reduced, the complexity of the algorithm is still exponential in the number of corresponding variables. The number of possible configurations that have to be checked by the argmax-operation is $\binom{M}{N}$ for an N-to-M mapping. Further reductions can be obtained by substituting the argmax-operation with an approximating search over possible variable configurations.

Returning to the tennis example (Fig. 3.9(a)), the approach of conditional buckets works as follows. Let us assume the same variable ordering as in section 3.3.3: $S,T,OP''_a,OP''_b,OP'_1,OP'_2,R,OP_1,OP_2,NP_1,NP_2,N_a,N_b$. Conditioning over S,T results in the following bucket allocation:

$$
\begin{aligned}
\mathcal{N}_b &: P(OP''{}_b|N_b = Lendl)\ P(N_b = Lendl)\\
\mathcal{N}_a &: P(OP''{}_a|N_a = Agenor)\ P(N_a = Agenor)\\
\mathcal{NP}_2 &: P(OP_2|OP'_2, NP_2 = false)\ P(NP_2 = false)\\
\mathcal{NP}_1 &: P(OP_1|OP'_1, NP_1 = true)\ P(NP_1 = true)\\
\mathcal{OP}_2 &:\\
\mathcal{OP}_1 &:\\
\mathcal{OP}'_2 &: \begin{cases} [\mathcal{OP}'_2|T = \tilde{o}p''_a]: & P(OP'_2|OP''_a, T = \tilde{o}p''_a)\\ [\mathcal{OP}'_2|T = \tilde{o}p''_b]: & P(OP'_2|OP''_b, T = \tilde{o}p''_b)\end{cases}\\
\mathcal{OP}''_1 &: \begin{cases} [\mathcal{OP}'_1|S = \tilde{o}p''_a]: & P(OP'_1|OP'_a, S = \tilde{o}p'_a)\\ [\mathcal{OP}'_1|S = \tilde{o}p'_b]: & P(OP'_1|OP'_b, S = \tilde{o}p''_b)\end{cases}\\
\mathcal{OP}''_b &:\\
\mathcal{OP}''_a &:\\
[\mathcal{S}, \mathcal{T}] &: P(T)\ P(S)
\end{aligned}
$$

The first six buckets are eliminated straightforward resulting in:

$$
\begin{aligned}
\mathcal{OP}'_2 &: \lambda_{OP_2}(OP'_2) \begin{cases} [\mathcal{OP}'_2|T = \tilde{o}p''_a]: & P(OP'_2|OP''_a, T = \tilde{o}p''_a)\\ [\mathcal{OP}'_2|T = \tilde{o}p''_b]: & P(OP'_2|OP''_b, T = \tilde{o}p''_b)\end{cases}\\
\mathcal{OP}'_1 &: \lambda_{OP_1}(OP'_1) \begin{cases} [\mathcal{OP}'_1|S = \tilde{o}p''_a]: & P(OP'_1|OP''_a, S = \tilde{o}p''_a)\\ [\mathcal{OP}'_1|S = \tilde{o}p''_b]: & P(OP'_1|OP''_b, S = \tilde{o}p''_b)\end{cases}\\
\mathcal{OP}''_b &: \lambda_{N_b}(OP''_b)\\
\mathcal{OP}''_a &: \lambda_{N_a}(OP''_a)\\
[\mathcal{S}, \mathcal{T}] &: P(T)\ P(S)
\end{aligned}
$$

The buckets $\mathcal{OP}'_2, \mathcal{OP}'_1$ generate conditional messages that are propagated to the buckets $\mathcal{OP}''_b, \mathcal{OP}''_a$:

$$
\begin{aligned}
\mathcal{OP}''_b &: \lambda_{N_b}(OP''_b) \begin{cases} [\mathcal{OP}''_b|T = \tilde{o}p''_b]: & \lambda_{OP'_2}(OP''_b, T = \tilde{o}p''_b)\\ [\mathcal{OP}''_b|T \neq \tilde{o}p''_b]: & \begin{cases}[\mathcal{OP}''_b|S = \tilde{o}p''_b]: & \lambda_{OP'_1}(OP''_b, S = \tilde{o}p''_b)\end{cases}\end{cases}\\
\mathcal{OP}''_a &: \lambda_{N_a}(OP''_a) \begin{cases} [\mathcal{OP}''_a|T = \tilde{o}p''_a]: & \lambda_{OP'_2}(OP''_a, T = \tilde{o}p''_a)\\ [\mathcal{OP}''_a|T \neq \tilde{o}p''_a]: & \begin{cases}[\mathcal{OP}''_a|S = \tilde{o}p''_a]: & \lambda_{OP'_1}(OP''_a, S = \tilde{o}p''_a)\end{cases}\end{cases}\\
[\mathcal{S}, \mathcal{T}] &: P(T)\ P(S)
\end{aligned}
$$

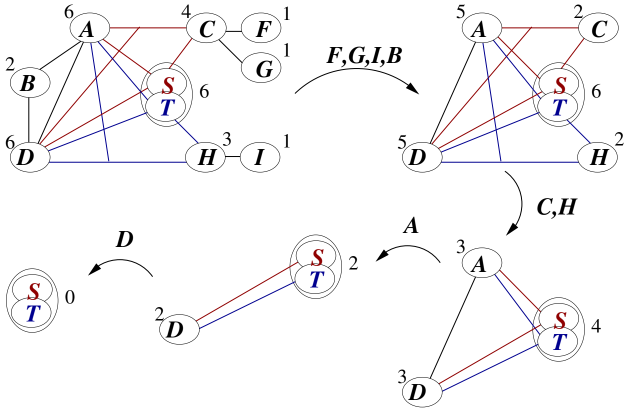


Figure 3.18: Generating the variable order $[S,T],D,A,H,C,B,I,G,F$ for the conditional bucket scheme. The small numbers denote the actual induced width.

Due to the exclusive constraint on possible assignments of S,T only two conditional messages are calculated for the buckets $\mathcal{OP}''_b$ or $\mathcal{OP}''_a$, respectively:

$$[\mathcal{S},\mathcal{T}] : P(T)\,P(S)$$

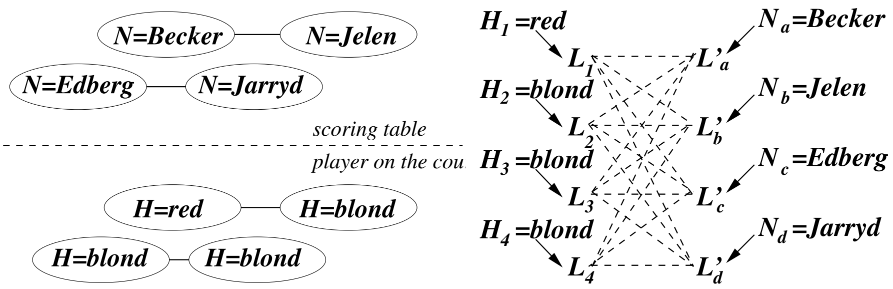
$$\begin{bmatrix} 0 & \lambda_{OP''_b}(T=\tilde{op}''_b)\,\lambda_{OP''_a}(S=\tilde{op}''_a, T \neq \tilde{op}''_a) \\ \lambda_{OP''_b}(S=\tilde{op}''_b, T \neq \tilde{op}''_b)\,\lambda_{OP''_a}(T=\tilde{op}''_a) & 0 \end{bmatrix}$$

This time, for any variable mapping separate messages are calculated that are then combined in the exclusive bucket. Therefore, the number of conditional messages is only $N \times M$. The conditional bucket scheme automatically contributes to such independency assumptions that are introduced by the structure of the Bayesian network. On the other hand it is general enough to solve more complex problems like the aforementioned example (Fig. 3.17(b)). There, a conditional probability table models a relevance relation between the corresponding variables A,D that violates the independency assumptions that exist in the tennis example.

The variable ordering for the conditional bucket scheme can be calculated similarly to the normal bucket scheme by successively selecting the node of the moral graph with the minimal induced width (Fig. 3.18). The combined exclusive bucket $[\mathcal{S},\mathcal{T}]$ is treated as one node in the moral graph. The CPTs $P(C|A,S=\tilde{a}),P(C|D,S=\tilde{d})$ induce a special treatment. For the node C it is treated like a unique CPT $P(C|X,S)$ where X depends on the value of S. Therefore, the edge in the moral graph is split into two edges to the nodes A and D. For the nodes A,D these conditioned CPTs must be considered separately.

3.4 Relation to Graph Matching

The modeling of the corresponding variables and the conditional bucket elimination scheme is tightly related to probabilistic graph matching. The two sets of corresponding variables denote the nodes of the two graphs that shall be matched. The selection



(a) Graph structure (b) Corresponding variables

Figure 3.19: A graph-matching example in the tennis domain. N denotes the names on the scoreboard, H the hair color, and L, L' the anticipated look of the players. The edges represent the relation that the names or players are members of the same team.

variables $S_1, \ldots, S_n \in \mathcal{W}$ realize the mapping function from the nodes of the one graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ into the other $\mathcal{H} = (\mathcal{W}, \mathcal{F})$. Edges between the nodes can be formulated by relations over the selection variables. For each edge $e_{ij} = \{v_i, v_j\} \in \mathcal{E}$ a relation R_{ij} is introduced with CPT:

$$P(R_{ij} = 1 | S_i = s_k, S_j = s_l) = \begin{cases} 1.0, & \text{if } \{s_k, s_l\} \in \mathcal{F} \\ 0.0, & \text{otherwise} \end{cases}$$

The search for a correct match is equal to finding a maximum a posteriori hypothesis of the selection variables:

$$(s_1^*, \ldots, s_n^*) = \operatorname*{argmax}_{s_1, \ldots, s_n,\, s_i \neq s_j, i \neq j} P(S_1, \ldots, S_n | \mathbf{e}) \prod_{\{v_i, v_j\} \in \mathcal{E}} P(R_{ij} = 1 | S_i, S_j)$$

For example, extending the tennis example to a double with four players on the scoreboard and four players on the court results in a graph matching problem (Fig. 3.19). The nodes of the first graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ are labeled with the names on the scoreboard $N \in \{\textit{Becker,Jelen,Edberg,Jarryd}\}$. The nodes of the other graph $\mathcal{H} = (\mathcal{W}, \mathcal{F})$ are labeled with the players' hair color $H \in \{\textit{blond, red}\}$. The edges represent the relation that the names or players are members of the same team. The names are included in the same scoreboard entry and the players are located on the same side of the tennis court. The node matching can be modeled by comparing the anticipated look of the players $L, L' \in \{\textit{German,Swedish,other}\}$. In the Bayesian network this leads to a 4-to-4 mapping of the corresponding L, L'-variables. The graph edges can be modeled by two CPTs $P(R_{12} = 1 | S_1, S_2), P(R_{34} = 1 | S_3, S_4)$, $S_1, \ldots, S_4 \in \{\tilde{l}'_a, \ldots, \tilde{l}'_d\}$ that are defined identically:

$$P(R_{ij} = 1 | s^{(i)}, s^{(j)}) = \begin{bmatrix} 0.0 & 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \end{bmatrix}$$

The graph match may be probabilistic due to different reasons. First, a node can be labeled with a probability distribution of discrete values instead of a single value. Secondly, there may an edge between two nodes in the graph $\mathcal{H}$ with a certain probability instead of only existing or not existing. A labeled edge can be modeled by extending the domain $\Omega_{R_{ij}} = \{0,1\}$ of R to the appropriate label set $\mathcal{L}_{ij}$ plus an extra label ε for *'no edge at all'*: $\Omega_{R_{ij}} = \mathcal{L}_{ij} \cup \{\varepsilon\}$. The graph matching then considers a probability distribution over this label set. Traditional graph matching algorithms [Mes95] assume an independent scoring of node matches and edge matches. The Bayesian network approach combined with the conditional bucket elimination method does not explicitly need these independency restrictions. The technique automatically adapts to such conditional independencies (cf. the Bayesian networks in Fig. 3.17(b) and Fig. 3.10(a)).

3.5 Applications of Bayesian Networks

The following subsections will present different examples of the application of Bayesian networks that have been proposed in the literature. The first one is a classic example from the medicine domain that combines conditional probability tables that have been estimated from examples and those tables that have been calculated from a computational model. The others are examples in the context of object recognition: modeling aspect hierarchies, attention, or arrangements of objects.

The MUNIN system

MUNIN (MUscle and Nerve Inference Network) is a causal network for the interpretation of electromyographic findings, that is the diagnosis of muscle and nerve diseases through analysis of bioelectrical signals from the muscle and nerve tissue [AWFA85]. It was one of the first Bayesian networks of non-trivial size for a realistic application. The structure of the network is shown in Fig. 3.20. The DIAGNOSIS node includes the finding of three different diseases, each with two to four states plus states for *Normal* and *Other*. The mediating nodes consist of eight pathophysiological nodes that describe the changes in a given muscle and one node for integrating three different findings. The nodes MU.LOSS, POSTSYN.NEU.MUSC.TRANS, PRESYN.NEU.MUSC.TRANS, and the integration node MUP.CONCLUSION have an additional state *Other*. Fifteen different findings are modeled by fifteen evidential nodes with two to seven states each.

Because the states of the three different diseases have been subsumed under the same node, multiple diseases cannot be considered by the Bayesian network. Another restriction is that only the findings from one single muscle can be inserted into the network.

As far as possible, the conditional probabilities have been calculated based on “deep knowledge”, i.e. physical or physiological models that describe the interrelation of the underlying pathophysiological interpretations of the statistical variables. Starting from a finding of MU.LOSS and MU.STRUCTURE, these are translated into variables that are compatible with the model. Then a new model variable is calculated utilizing the “deep knowledge”. After that, the new variable is translated into the statistical variable,

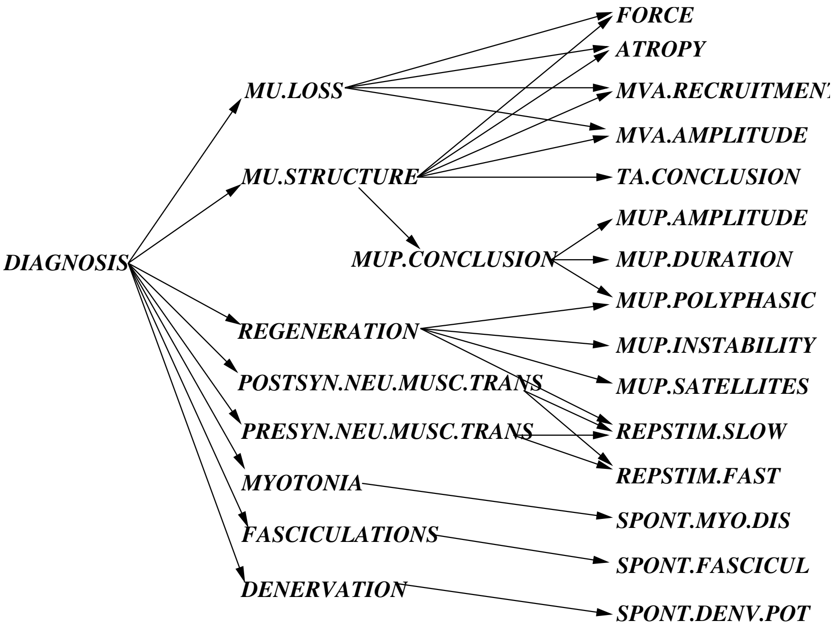


Figure 3.20: Bayesian network structure of MUNIN [AWFA85].

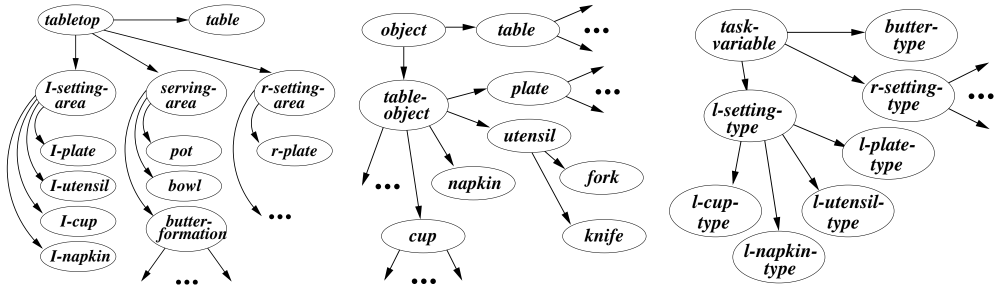
e.g. ATROPHY. For findings with continuous outcomes the probabilities are replaced by probability distributions. The *Other* states of the variables have relatively even distributions. Thereby, these states obtain support for conflicting cases that are not covered by the modeled states. The authors mention two different interpretations of *Other* states. If the findings entered into the network faithfully represent the state of the muscle, a “hole” in the knowledge of the network has been discovered. Alternatively, erroneous findings may have been entered.

The network is verified by two different experiments. First, the network is used in top-down manner by generating expectations of findings corresponding to a single disorder. Secondly, typical findings of different diseases are entered, and the diagnosis is calculated bottom-up.

The TEA-1 composite network

Raymond D. Rimey and Christopher M. Brown present the TEA-1 system that employs Bayesian networks and decision theory in order to realize systems capable of ***hypothesis-driven sufficient vision*** [RB94]. The term ***sufficient vision*** introduces a paradigm in which structured scenes are processed selectively, i.e. using only some of the vision modules, analyzing only small areas of an image, interpreting only sufficient details, using knowledge earlier in the process, actively controlling the sensor, and solving specific visual tasks instead of reconstructing all objects in the entire image.

The TEA-1 system solves visual tasks by gathering scene evidence using visual operation. In [RB94] the authors present a system for classifying *dinner-table-scenes*. The



(a) Part-of net/area net (b) Is-a net (c) Task net

Figure 3.21: Bayesian networks of the TEA-1 system [RB94].

scenes are structured in that they consist of an arrangement of objects with a recursively defined spatial grouping. The table scene is structured by a setting area for each person and a serving area. These are structured themselves, e.g. the setting area consists of a plate, a knife, a folk, and a cup. Rimey and Brown introduce the term ***T-world*** for such domains.

The knowledge in the TEA-1 system is specified in four different Bayesian networks. The recursive structure is represented in the *part-of net* (Fig. 3.21(a)). Geometric relations are modeled in the *expected area net* that has the same structure as the *part-of net*. A node in the first network identifies a particular object, and the corresponding node in the *expected area net* identifies the area in the scene in which this object is expected to be located. The domain of the random variables in this network are the positions on a discrete two-dimensional grid. Conditional probability tables between variables A, B are defined with regard to a simplified distribution called ***relational map*** $R_{B|A}(x,y)$. A relational map assumes that object A has unity dimensions and is located at the origin. The conditional probabilities are obtained from a function f that is given by a knowledge engineer:

$$P(p_B|p_A) = f(p_B; R_{B|A}, p_A, h_A, w_A), \quad \text{where } h_A, w_A \text{ are the height and the width of object } A$$

The classification of each object in the scene is represented by an *is-a net* (Fig. 3.21(b)). It models a taxonomic hierarchy of mutually exclusive subset relationships in the domain. Each node of the *part-of net* is associated with an *is-a net*.

The task-specific knowledge is separated from the domain knowledge described so far. It is encoded in a *task net* (Fig. 3.21(c)) which specifies what objects and object property values are expected for each possible outcome of the task variable.

The four separate Bayesian networks are linked within the *composite net*. The propagation of evidences in this net is realized as follows:

1. Propagate belief in each of the separate nets in the composite net except for the task net.

2. Construct *packages* of *BEL* values from the other nets for transfer to the task net. These packages define the evidences that are attached to the nodes in the task net.

3. Propagate belief in the task net.

TEA-1 currently uses combined beliefs about the presence of an object in the scene and the detailed classification result ω_i. For the *l-utensil* node the package is:

$$
\begin{aligned}
&(\alpha\beta_{fork}, \alpha\beta_{knife}, \alpha\beta_{spoon}, 1-\alpha(\beta_{fork}+\beta_{knife}+\beta_{spoon}))\\
&\quad \text{where } \alpha = BEL(present) \quad \text{— from the part-of net,}\\
&\qquad\qquad \beta_i = BEL(\omega_i), \quad i \in \{fork, knife, spoon\} \quad \text{— from the is-a net.}
\end{aligned}
$$

The next action of the system is selected on the basis of the propagated beliefs of the networks. Either *visual actions* or *camera movement actions* can be performed. Therefore, the problem of *which evidence to get next* is extended by the decision *where to look for evidence*. Each kind of object usually has several actions associated with it. In the table-setting domain, TEA-1 currently has 21 visual actions related to seven kinds of objects. Any action has a precondition that has to be fulfilled before the action is executed. For example, the *per-detect-plate* action can be performed if the plate's location is not yet known, if the expected location of the plate is within the visual field for the current camera position, and if the action has not been executed previously. The decision of the best action to be performed is based on the specific costs of an action and the expected effort of an action. The latter is measured by the ***expected value of sample information*** (*EVSI*). The expected value of the task decision d_i, here if the table setting is *fancy* (d_0) or *not-fancy* (d_1), is defined as

$$
\begin{aligned}
&EV(d_i) = \sum_{j=0}^{1} V(d_i,t_j)\, P(t_j)\\
&\qquad \text{where } V(d_i,t_j) = \begin{cases} 1000, & \text{if } i = j \\ -1000, & \text{if } i \neq j \end{cases} \quad \text{is a the payoff function.}
\end{aligned}
$$

Here, $P(t_j)$ is the actual belief of the task node if the table setting is *fancy* or *not-fancy*. $EV_0 = \max_i EV(d_i)$ is the payoff value of the optimal decision. The expected payoff value EV_e after performing the action can be defined by means of the piece of evidence e that may be extracted by the action:

$$
EV_e = \sum_{k=0}^{n_e} \left[\max_i \left\{ \sum_{j=0}^{1} V(d_i,t_j)\, P(t_j|e_k)] \right\} \right] P(e_k)
$$

$$
\text{where} \quad n_e \text{ is the number of possible values of } e.
$$

The expected value of the sample information is then given by the difference between the expected value of the task decision before and after the action is executed:

$$
EVSI(e) = EV_e - EV_0
$$

The control loop of TEA-1 does not only consider the next action but also sequences of possible actions when deciding which action is to be performed next. The payoff function weights the camera movements against the visual actions and provides a threshold criterion for succeeding.

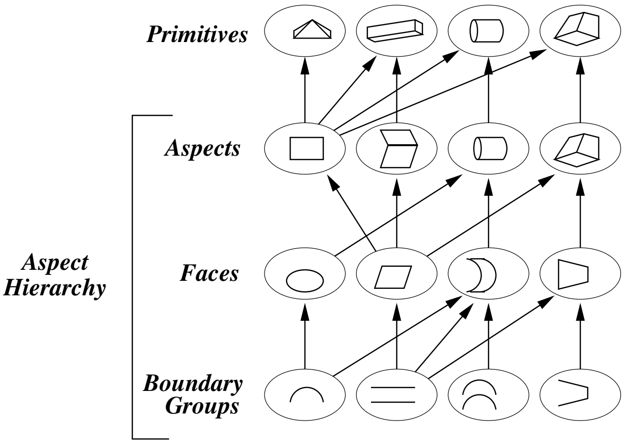


Figure 3.22: The aspect hierarchy for recognizing primitives [DPR92].

The probability values of the different Bayesian networks are specified by a human who is familiar with the application domain and task. It turned out that the general behavior of the system is relatively insensitive to variations in the values of the supplied probabilities.

The OPTICA aspect hierarchy

Dickinson et al. present the OPTICA system that employs Bayesian networks in a 3-d shape recovery approach [DPR92]. An object recognition task is divided into the finding of 3-d volumetric primitives that could be used for object indexing and the recognition of complex objects that can be constructed by combining these primitives. The idea of this approach is to extract a small finite set of powerful indexing primitives in order to access a large, may be infinite, object databases. The computational burden is, therefore, partially shifted from top-down verification of simple 2-d features towards the bottom-up extraction and grouping of 2-d features into volumetric primitives.

The Bayesian network approach is used to constrain the search of the grouping process. Given an arbitrary set of 3-d primitives, an aspect hierarchy is constructed that realizes a probabilistic mapping from 2-d features to these volumetric primitives. In order to account for occlusions in the scene, the aspect hierarchy is organized in levels of different complexity (Fig. 3.22):

- ***Aspects*** constitute the top level of the hierarchy. They represent the set of topologically distinct views of the primitives.

- ***Faces*** correspond to the primitive surfaces. Combinations of them define the aspects.

- ***Boundary groups*** represent all subsets of lines and curves comprising the faces. They define the lowest level of the aspect hierarchy.

All elements of the aspect hierarchy *qualitatively* represent geometric elements of the image. Boundary groups and faces are defined by a graph of qualitative relationships (*intersection, parallelism, symmetry*) among qualitatively described contours, e.g. two parallel lines of equal length is a boundary group that is a subgraph of several face graphs. Aspects denote graphs in which nodes represent faces and edges represent face adjacencies.

The hierarchy consists of 37 different aspects, 16 faces, and 31 boundary groups. The ambiguities in the mapping from a lower level to the next higher level are captured by conditional probability tables:

$$P(\mathit{primitive}|\mathit{aspect}), P(\mathit{aspect}|\mathit{face}), P(\mathit{face}|\mathit{boundary\text{-}group}).$$

The CPTs are estimated from simulated data. The primitive volumes are rotated around their internal axes and projected onto the image plane. The appearance of each feature and its parent is noted and counted. A priori probabilities of occurrence or orientation of primitives can be considered during the simulation process.

The Bayesian network that is defined by the three conditional probability tables is exploited during shape recovery in the following way: First, a contour graph is calculated in which nodes denote curvature discontinuities or junctions of contours and in which edges are the actual bounding face contours. From this graph, closed image faces are extracted. If the image face exactly matches a face of the aspect hierarchy this is directly used as an evidence:

$$\underline{e}_{\mathit{face}}(i) = \begin{cases} 1.0, & \text{if } \mathit{face}\ i \text{ has exactly been matched} \\ 0.0, & \text{otherwise.} \end{cases}$$

If there is no exact match the Bayesian network is instantiated on a lower level using the boundary groups found for the image face:

$$\underline{e}_{\mathit{boundary\text{-}group}}(i) = \begin{cases} 1.0, & \text{if } \mathit{boundary\ group}\ i \text{ has been matched} \\ 0.0, & \text{otherwise} \end{cases}$$

resulting in a probability distribution of face labels. Then the probability of the most probable explanation of an *aspect* hypotheses a is calculated that might encompass that face. Either the evidence $\mathbf{e} = \underline{e}_{\mathit{face}}$ or $\mathbf{e} = \underline{e}_{\mathit{boundary\text{-}group}}$ is given:

$$P(\mathit{aspect} = a|\mathbf{e}) = \max_{b,f} P(\mathit{aspect} = a, \mathit{face} = f, \mathit{boundary\text{-}group} = b|\mathbf{e})$$

This probability is used in order to constrain the search for an aspect instantiation, i.e. the finding of a graph match of an aspect from the aspect hierarchy with a subset of image faces. In the same way, the search process for the primitives can be constrained by the aspects found:

$$P(\mathit{primitive} = p|\underline{e}_{\mathit{aspect}})$$

During the face labeling process, the aspect hierarchy and the conditional probabilities defined in it can be additionally exploited in order to get rid of segmentation errors. For this purpose, the authors present a model-based region merging algorithm. Starting with an over-segmentation, the algorithm merges two faces if the probability of a face label can be increased.

Recognizing collaborative multi-agent activities

Steven S. Intille presents a multi-agent recognition scheme using Bayesian networks [Int99]. American football games are classified according to different team actions. These collaborative activities are described by low-order spatial and temporal relationships between players and single player goals. The recognition scheme is organized in three different levels. First, *perceptual features* are calculated from player and ball trajectories like velocity, curvature, relative orientation of objects, distance between objects, entering of special regions, path interception estimation, etc. Secondly, these features are combined in *visual networks* that provide likelihoods for agent goals, like *object1 catches a pass*. On the third level, these likelihoods are used as evidences for *multi-agent networks*. For each specifiable collaborative team action class there is one multi-agent network. The nodes of these Bayesian networks model compound goals, binary temporal relationships, observation of goals, or logical relationships between goals or compound goals. Thereby, the integration of evidence over time and the evaluation of temporal relationships is implicitly encoded in the nodes of the network, not within the linking structure. The correspondence problem of assigning a model object to a visually observed object is not solved in the Bayesian network. Instead, the evaluation of the Bayesian network finds a consistent interpretation for a given object assignment. The object assignment is calculated using some additional preference information and a rule-based algorithm.

In summary, three modeling principles can be identified in the work of Steven S. Intille. First, the usage of decoupled visual and multi-agent networks in order to modularize the modeling task. Secondly, the encoding of temporal relationships as single evidence nodes. Thirdly, the external solution of the correspondence problem by defining the multi-agent networks as functions of a specific object assignment and using an external heuristic search.

3.6 Bayesian networks for integration of speech and images

At the beginning of this chapter, Bayesian networks were introduced as an inference calculus for uncertain reasoning. Bayesian networks are well founded in probability theory, provide intuitive notions for modeling relevance relationships between domain variables, and offer causal and diagnostic ways of reasoning.

3.6.1 Modeling principles

Chapter 2 discussed several aspects of speech and image integration that require a treatment of different kinds of uncertainty: noisy input data; propagated errors in abstraction hierarchies; lexical, syntactic, semantic, and pragmatic ambiguities in verbal descriptions; positional and orientational uncertainty in spatial reasoning; unknown selections of a reference frame; implicit contexts. In this work, the position is taken up that all different kinds of uncertainties can be modeled by the same probabilistic formalism, i.e they are assumed to combine like frequencies.

The design task in Bayesian networks consists of selecting an appropriate network structure and determining the numbers of the conditional probability tables. Applications of Bayesian networks reported in the literature give hints on how to design the structure and get the numbers.

- The vocabulary problem in human-computer interaction, unknown objects, and erroneous classification results are all related to out-of-scope events. In the MUNIN-system out-of-scope events are modeled by *Other* states in the Bayesian network that nearly have an even distribution. By this means, unexpected or erroneous evidence can be detected.

- Dickinson et al. demonstrate how to cope with occlusions and segmentation errors in object recognition. They introduce an aspect hierarchy that is able to consider evidence on different levels of abstraction and complexity. Missing information is hypothesized, and the merging operation is controlled by exploiting the statistical knowledge modeled in the Bayesian network.

- Common to the TEA-1 network of Rimey and Brown and that of Intille is the modularization into more than one network. The calculated beliefs of the first group of networks is used as evidence in a second network that provides the final classification. In TEA-1 this is the task net, in the work of Intille it is the multi-agent network. On the one hand, the modeling and propagation is simplified by using several small networks. On the other hand, flexibility that is provided by Bayesian networks is partially lost, or has to be realized externally to the formalism, e.g. causal and diagnostic inference.

- The last chapter identified the solution of the correspondence problem as the central task in speech and image integration. All different kinds of uncertainties contribute to the referential uncertainty, i.e. which object is denoted in a scene. One aspect of this is that object class and shape information has to be coupled with spatial information. In TEA-1 this correspondence is fixed in the part-of and expected-area nets that have the same structure and corresponding nodes. In the work of Intille the correspondence of model players and player trajectories in the image is calculated externally.

- The numbers of the conditional probabilities are either set by hand (TEA-1, Intille), calculated from computational models (MUNIN, TEA-1), or estimated using experimental (MUNIN) or simulated data (OPTICA).

This chapter presented a solution of the correspondence problem in the language of Bayesian networks. Thereby, modularly defined partial networks can be linked to a homogenous network that can be evaluated by a one pass algorithm. The inherent flexibility of Bayesian networks is preserved. The correspondence problem is translated into the finding of a mapping of corresponding variables that is controlled by a set of exclusive selection variables. The general structure of the proposed Bayesian network is shown in Fig. 3.23. Each ***v-object*** in the visually observable scene and each ***s-object***

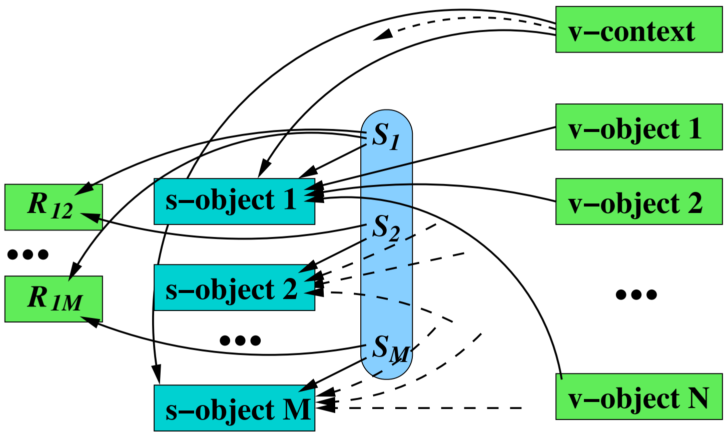


Figure 3.23: General structure of a Bayesian network for speech and image integration. $S_1, \ldots, S_M$ are exclusive selection variables that control the mapping between speech objects and visual objects. Filled boxes denote Bayesian subnetworks.

mentioned in a spoken utterance is described by a separate subnetwork. In both kinds of subnets corresponding variables are defined and linked by exclusive selection variables $S_i \in \{\tilde{v}_1, \ldots, \tilde{v}_N\}, i = 1, \ldots, M$. Additionally, the Bayesian subnets of the s-objects may be influenced by a visual context variable that provides a more general scene context that might induce a shift in the meanings of words. Verbally mentioned relations between objects are modeled by probabilistic relations between selection variables. The conditional probabilities of such relations are determined by visual object properties.

3.6.2 Inference methods

Bayesian networks are intensional models, i.e. the inference algorithm is separated from the modeling task. Thus, different inference tasks like belief updating, finding the most probable explanation, or finding the maximum a posteriori hypothesis can be realized for the same Bayesian network representation.

Nevertheless, the complexity of the inference algorithm strongly depends on the structure of the network. For singly connected networks there is the recursive algorithm of Pearl that is linear in the number of nodes. The evaluation of networks that include undirected circles is NP-hard in general, but as long as the number of loops is controllable, the inference task is tractable for generalized algorithms. However, in such a case the complexity of the inference task depends on how the control is realized, which variable is selected for conditioning, how the moral graph is triangulated for the construction of the junction tree, or which order of variables is selected in the bucket scheme.

The proposed Bayesian network structure that models the mapping of corresponding variables is best suited for a combined application of the conditioning technique and the bucket elimination scheme. A realization of such a combined strategy can be formulated by extending the bucket scheme with conditional and exclusive buckets. In this chapter this idea has been worked out and demonstrated by different examples. It turned out that such a technique is tightly related to probabilistic graph matching, but automatically adapts to introduced independencies instead of being implicitly coded in the matching algorithm.

3.6.3 An application to human-computer interaction

The next chapter will apply the introduced principles and techniques to a human-computer interface of a robot that is acting in the real world. The program module proposed in this thesis takes inputs from object recognition and speech understanding modules and solves the correspondence problem in a tight coupling with a dialog component. It will be shown that additional inferences are possible that detect and correct erroneous recognition results, complete verbal descriptions, and classify unknown objects.

Chapter 4

Modeling

4.1 Scenario and Domain Description

The system described in this chapter was realized in the context of the Collaborative Research Center 360 ”situated artificial communicators”. This long term project aims at the development of a robot constructor which can be instructed by speech and gestures in the most natural way. This chapter will concentrate on the integration of spoken instructions and the visually observed scene.

While the communication between the human instructor and the system is constrained as little as possible, the domain setting is rather restricted. A collection of 23 different elementary baufix®[1] components[2] is lying on a table (Fig. 4.1). Most of these ***elementary objects*** have characteristic colors, e.g. the length of a bolt is coded by the color of its head. Nevertheless, the colors do not uniquely define 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 on it, screw or plug them together, and can put them down again. The actions that shall be performed by the robot are verbally specified by a human. In contrast to the system that has specific domain knowledge about the baufix®world, 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 baufix®terms:[3]

- *”gib mir die kleine runde, den kleinen runden Holzring ;”*
  (Give me the small round, the small round wooden-made ring.)

- *”ich möchte den großen grünen Würfel wo auf jeder Seite ein Loch ist ;”*
  (I would like the big green cube with a hole on each side.)

[1] baufix®is a wooden toy construction kit from the company Lorenz.

[2] The elementary baufix®objects are listed in appendix A.

[3] The following examples are taken from the data set of experiment 5 that will be described in the next section. The spoken instructions have been transcribed.

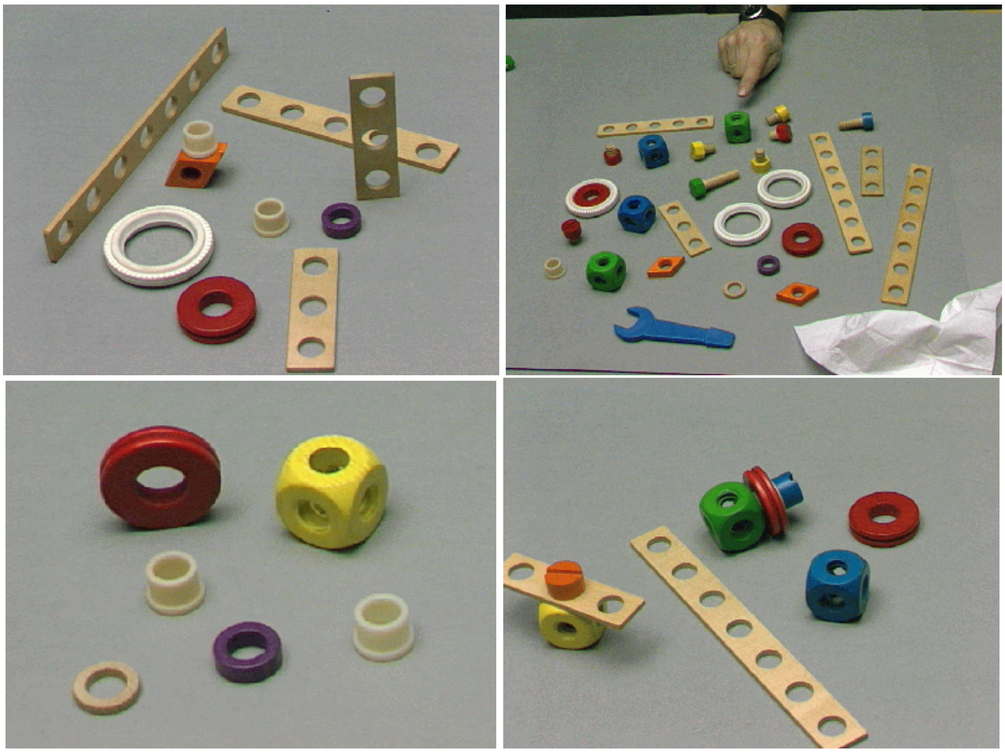


Figure 4.1: Table scenes in the baufix®domain.

- *”das ist ein äh flacher, äh eine flache Scheibe mit einem Loch ;”*
  (That is a er flat, er a flat disc with a hole.)
- *”den hellen Ring neben dem blauen Ring ;”*
  (The bright ring next to the blue ring.)
- *”die grüne eckige Mutter ;”*
  (The green angular nut.)
- *”gib mir die kleine lila Schra/[ube] Scheibe ;”*
  (Give me the small purple bo/[lt] disc.)
- *”Dreilochleiste zwischen der Fünflochleiste und der [Siebenlochleiste] ;”*
  (Three-holed bar between the five-holed bar and the [seven-holed bar].)
- *”ich möchte eine Dreierleiste und zwar liegt die ganz rechts ;”*
  (I would like the three-holed bar and it lies on the very right.)

Typically, the human instructor has a ***target object*** in mind that shall be constructed, e.g. a toy-plane or a toy-truck. Such a construction consists of several assembly steps and subgoals. For example, before the toy airplane can be assembled, a propeller, an engine block, a cockpit, a tail unit, and a landing gear must be constructed first. The instructor will use this terminology during the construction process, e.g. *”take the propeller and screw it onto the engine block”*. Such terms typically denote ***complex objects*** that consist of several connected elementary objects and have been constructed during the assembly process. The system has no pre-defined semantics for words like *propeller* or *engine block* and must treat them as ***unknown words*** (cf. 1).

4.2 Experimental Data

In order to capture the language use of unexperienced human instructors, a series of experiments were conducted [BJPW95, SSP00, Vor01a]:

Experiment 1 (Baufix) *27 subjects were asked to verbally describe 34 elementary* baufix® *objects separately (total number of words: 13,845). The scene context varied between isolated objects and context objects.*

Experiment 2 (Human-Human) *This sample consists of 18 human-human dialogs (total number of words: 13,726). One subject was told to be the instructor, the other one was told to be the constructor. The aim of the dialog was the construction of a toy airplane.*

Experiment 3 (Human-Machine) *The human-machine communication was simulated by a Wizard-of-Oz scenario. 34 subjects took the role of the instructor. The constructor part was taken by the simulated machine. 22 construction dialogs were recorded (total number of words: 32,450).*

From these three samples frequently named color, shape, and size adjectives were extracted that are presented in Fig. 4.2 [Soc97, SSP00]:

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)*	rautenförmig *(diamond-shaped)*	breit *(large, wide)*
hell *(light)*	länglich *(elongated)*	hoch *(high)*
orange *(orange)*	dick *(thick)*	mittellang *(medium-long)*
lila *(violet)*	schmal *(narrow)*	mittelgroß *(medium-sized)*
holzfarben *(wooden)*	dünn *(thin)*	eckig *(angular)*

Figure 4.2: Frequently named color, shape, and size adjectives.

Experiment 4 (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* baufix® *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.*

In the following, only the data from German questionnaires were used because the domain language is German and the influences of foreign languages on the modeling should be eliminated. 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 has been performed by Constanze Vorwerg [Vor01a, Vor01b]. She extracted the following results:

1. All attributes except 'rund' *(round)* depend on context. But the context only partially determines the selection of it. Three-holed bars are less frequently named 'mittellang' *(medium-long)* if a five- and a seven-holed bar are context objects. But the frequency of the alternative selection 'kurz' *(short)* does not exceed that of 'mittellang' *(medium-long)*. 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 baufix®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)*.

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 directly extract the applicability of such attributes from image features. The solution that has been applied in this thesis is to use object class statistics. This approach was already proposed in previous work of Gudrun Socher [SSP00].

Experiment 5 (Select-Obj) *10 subjects verbally named objects in 10 different* baufix®*scenes that were presented on a computer screen. The scenes contained between 5 and 30 elementary* baufix®*objects. One object was marked by an arrow and the subject was supposed to give an instruction like* "Take the yellow cube.". *453 verbal object descriptions were collected (total number of words: 2394).*

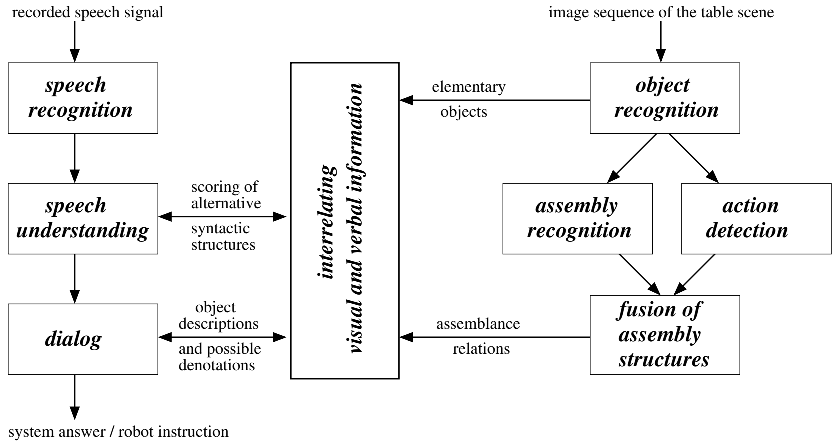

recorded speech signal
image sequence of the table scene
speech recognition
speech understanding
dialog
interrelating visual and verbal information
object recognition
assembly recognition
action detection
fusion of assembly structures
elementary objects
scoring of alternative syntactic structures
object descriptions and possible denotations
assemblance relations
system answer / robot instruction

Figure 4.3: Interaction of sysem components.

Experiment 6 (Select-Rel) *The experimental setting was equal to the experiment described before. Now, 6 subjects were explicitly told to name objects in 6 different scenes by using a spatial relation like* ”take the red object behind the cube.” *174 verbal object descriptions were collected (total number of words: 1396).*

The data of the last two experiments define the evaluation sets for the system. In the first experiment, objects are mostly described by adjectives and locative prepositions. In the second one, spatial relations were used.

4.3 The General System Architecture

The work proposed in this thesis is embedded in the work of other researchers in the Collaborative Research Center 360. The programming modules provided by this thesis are parts of a bigger demo system that is called ***artificial communicator*** (cf. Sec. 1) and shall assist a human in a construction task [BFF$^+$01]. In this section the general architecture of this system will be outlined.

The ***artificial communicator*** is divided into two tracks that run in parallel (Fig. 4.3). In the first track, the speech signal is recorded by a microphone. The speech recognizer transforms the signal into a structured[4] word sequence. From this sequence the speech understanding module reconstructs the meaning of the sentence considering different parsing possibilities. The most likely sentence structure is then passed to the dialog component which joins the meaning of the sentence with information from previous dialog steps. Interaction with visual information takes place in two different processing steps. When selecting the most likely parsing possibility, visual information is one scoring criterion. Before the interpretation of an utterance is passed to a robotic component or a system answer is returned to the speaker, verbal object descriptions are expanded

[4] During the recognition process domain specific grammar rules are applied to the recognized word sequence.

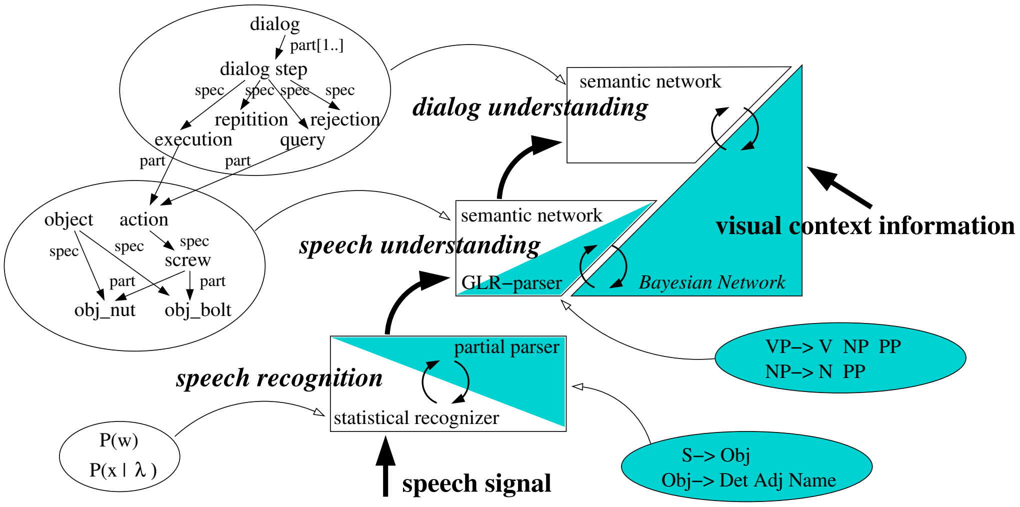


Figure 4.4: Speech understanding and dialog components.

by establishing referential links to visual object recognition results. Using these links, interactions like the correction or extension of an interpretation are performed.

In the second track, image sequences of the table scene are analyzed. First, ***elementary objects*** are recognized and coarsely reconstructed by stereo-matching. An assembly recognizer applies a syntactic approach in order to extract the structure of elementary object clusters, i.e. which objects are connected using which ports. By this means, ***complex objects*** are recognized. Simultanously, the construction process of complex objects is observed by an action detection algorithm that registers appearance and disappearance of objects. From this information the resulting assembly structure of an action is anticipated and fused with the result of the assembly recognizer when the complex object is put down into the scene.

The verbal and visual information streams are related by the integration component proposed in this thesis. The basic units of this integration process are elementary objects, named complex objects and relations between them. Spatial relations between an intended object and a reference object are often used to specify the location of an object. Assembly structures introduce two kinds of relations. First, if two elementary parts are assembled they have a *connected* relation. Secondly, the assembly structures define complex objects that have *part-of* relations to subassemblies and elementary objects.

The following two sections will present in more detail what is represented in an object hypothesis and in a verbal object description.

4.3.1 The speech understanding and dialog components

Fig. 4.4 shows the system from the speech processing perspective. The speech recognition, speech understanding and dialog components are not part of this thesis. The statistical speech recognizer has been developed and implemented by Gernot A. Fink [Fin99]. It is tightly coupled with a parsing component that constrains possible word sequences in a loose fashion, i.e. ungrammatical word sequences are still recognized [WFS98]. The speech understanding and dialog parts have been realized by Hans Brandt-Pook [BPFWS99, BP99]. Both parts are based on the semantic network formalism ERNEST [KNPS93]. The speech understanding component is coupled with a generalized LR-parser that has been developed and implemented by Susanne Kronenberg [KK99, Kro01]. It rearranges ***extrapositions***, i.e. corrections, completions, or complete constituents that are added by the speaker after the syntactical end of a sentence.

Speech recognition

The speech signal that is recorded by the wireless microphone is passed to the speech recognizer. Its core technologies are Hidden-Markov-Models (HMMs) (cf. e.g. [RJ93]) for describing acoustic events and statistical language models for providing estimates about word sequences likely to occur in the given domain. However, statistical language models, like n-grams, only provide useful restriction if sufficient training material is available. In artificial domains like the baufix®construction scenario this is typically not the case. Therefore, a partial parser provides additional restrictions on word sequences for the recognition process. The grammar that is utilized by the parsing process is given by a knowledge engineer. It declaratively describes expectations of possible word sequences that might be used by the speaker. Instead of modeling sentences, only important semantic parts of speech, like object descriptions, are specified. For example, if a sentence like *”Please, give me the <–> blue one which is, er, behind this long wooden stick”* is recognized, the structure in Fig. 4.5 is passed to the speech understanding component.

Speech understanding and dialog

Some semantic parts of speech like *”with the yellow bolt”* may be syntactically ambiguous. Either it modifies the verb of the sentence *”screw . . . with the yellow bolt”* and must

Please		
(*ACTION:* give)		
me		
(*OBJECT:*	(*ART:* the)	
	(*OBJ_ADJ: blue*))	
one which is er		
(*REF_OBJECT:*	(*SPATIAL_REL:* behind)	
	(*OBJECT:*	(*ART:*) this
		(*OBJ_ADJ:* long)
		(*OBJ_ADJ:* wooden)
		(*OBJ_NOUN:* stick)))

Figure 4.5: The structured word sequence that is passed from the speech recognizer to the understanding component for the sentence *”Please, give me the <–> blue one which is, er, behind this long wooden stick”*.

be interpreted as an instrument or it modifies the noun phrase *”… the cube with the yellow bolt”*. Another case of ambiguity is distinguishing between verbal corrections and and verbal extensions. Does the sentence *”take the long bolt <–> the blue one”* denote a *’long blue bolt’* or only a *’blue bolt’*? Often, such an ambiguity can be resolved by considering the visual context of the scene. Is there a cube on the table that is connected with a yellow bolt? Is there a blue bolt in the scene that is long? The ambiguity of such sentences results in two different representations that must be weighted by a component that integrates verbal and visual information:

a)	*OBJECT:*	*OBJ_NOUN:*	bolt
		ATTRIBUTES:	long
b)	*OBJECT:*	*OBJ_NOUN:*	bolt
		ATTRIBUTES:	long, blue

---

a)	*OBJECT:*	*OBJ_NOUN:*	cube	
		CONNECT_OBJECT:	*OBJ_NOUN:*	bolt
			ATTRIBUTES:	yellow
b)	*OBJECT:*	*OBJ_NOUN:*	cube	
	OBJECT:	*OBJ_NOUN:*	bolt	
		ATTRIBUTES:	yellow	

This is the first kind of query type that has to be answered by the Bayesian network proposed in this thesis:

Task 1 (*mp-interp*) *Which one of two or N possible interpretations of a verbal object description is most probable with regard to the visually observable scene?*

The answer is used as a scoring of alternative interpretations that are examined by a generalized LR-parser.

The semantic network transforms the parsing result of the generalized LR-parser into feature structures that are joined with those from earlier dialog steps. Then a second type of query is used in order to link the feature structures of verbal object descriptions to object instantiations in the visually observable scene:

Task 2 (*mp-objs*) *Which objects in the scene are most probably denoted by the interpretation of a verbal object description?*

Optionally, the context of scene objects can be restricted by the denotations from the interpretation of a previous verbal statement in the dialog. For example, the short dialog

USER: *”Take the bar.”*
SYSTEM: ”I have found two bars. Which one should I take?”
USER: *”The long one”*

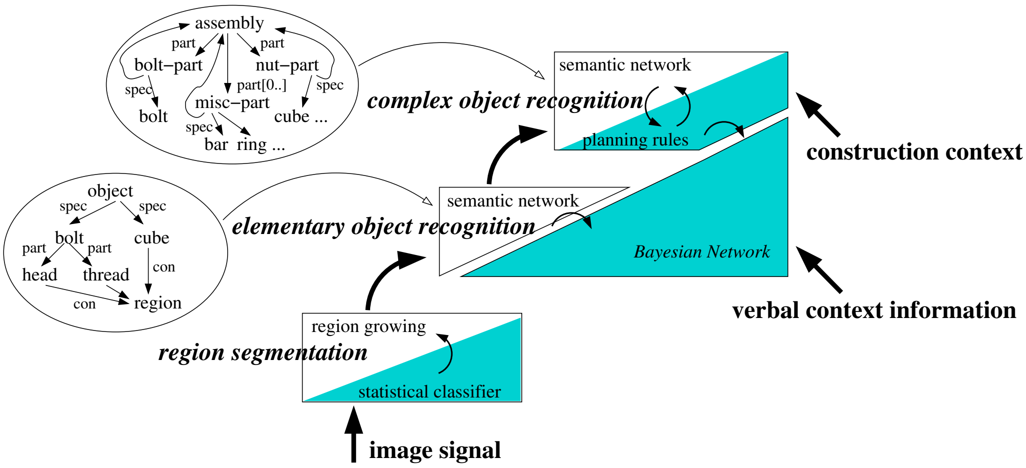


Figure 4.6: Recognition of elementary and complex objects

results in the queries:

OBJECT:	*OBJ_NOUN:*	bar

OBJECT:	*OBJ_NOUN:*	bar
	ATTRIBUTES:	long
	OBJ_RESTR:	`obj-5, obj-7`

where `obj-5, obj-7` are the selected scene objects from the first dialog step.

4.3.2 The object recognition component

The object recognition components of the system have mainly been realized by Franz Kummert, Elke Braun, Christian Bauckhage, and Jannik Fritsch [KFSB98, BKS98, BFKS99]. This section will describe those aspects of their work which are relevant for the understanding of this thesis. The elementary object recognizer integrates different processing cues like holistic hypotheses, regions, and contours. For the evaluation results presented in this thesis a simplified version is used that is only based on a single region cue.

The object recognition starts with a YUV-image. Each pixel is classified into one of the eight baufix®colors *red, orange, yellow, blue, green, violet, wooden, or ivory* plus one color for shadow and one for the background. A region-growing algorithm merges the classified pixels into regions of homogenous colors. A set of features (classified color, region size, eccentricity, compactness, etc.) that are calculated for every region are used as indexing primitives in the object database. Eleven different types of ***elementary objects*** plus one class for unknown objects (*3,5,7-holed-bar, cube, rhomb-nut, rim, tire, socket, flat-washer, thick-washer, bolt, undef*) are distinguished by a structural verification, that

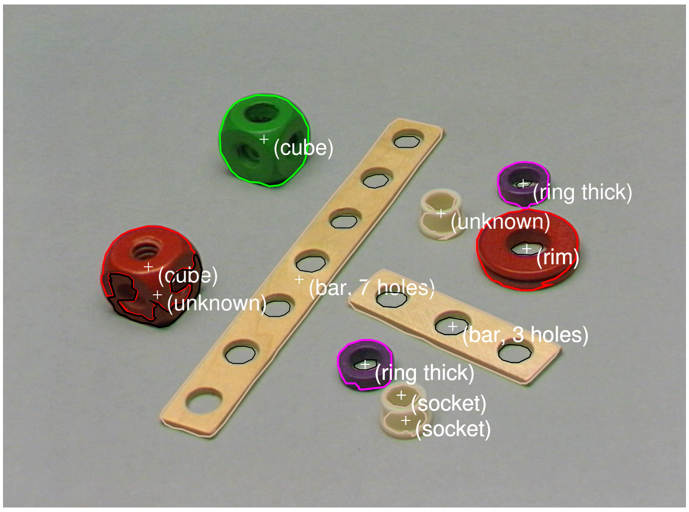


Figure 4.7: Recognition results of a typical scene in the baufix® scenario. One socket is incorrectly classified as *unknown*. The other one has been broken down into two sockets. Parts of the red cube are labeled as shadow.

is realized in a semantic network. For example, colored bolts consist of one optional wooden thread and a colored head. Red rims can be distinguished from red cubes by checking the existence of a hole in the center of the object region, etc. Fig. 4.7 shows an exemplary recognition result of a baufix® scene. If some of the indexing or structural criteria are not fulfilled due to the special aspect of an elementary object, an under- or over-segmentation, or an occlusion, either erroneous classification results occur, or the region is attached with an *undef* label. In such a case, the verbal context can be used in order extend the indexing primitive of the object recognition component:

Task 3 (*mp-class*) *Which is the most probable object class of an elementary object in the scene given the classification results of the vision component and a verbal description of it?*

In the next processing step of the vision component, assembly knowledge is used in order to recognize complex objects that have been constructed from elementary objects. Two independent tracks have been realized:

- The first track employs syntactical knowledge – coded in a *semantic network* – about how elementary objects can be connected. Each object cluster that has been detected in the visual scene is parsed and its structure is thereby extracted. For example, the structural analysis may be started with a *red bolt*. An adjacent object region is labeled *3-holed-bar*. It is classified as a *misc-part* and is assumed to be plugged onto the thread of the bolt. The next adjacent object is a *green cube* that can function as a *nut-part*. Consequently, it is assumed to fix the bar on the thread of the bolt. These three elementary object define a valid assembly that now itself is checked if it is the bolt-, misc-, or nut-part of another assembly.

- The second track observes the construction process of a complex object by detecting disappearing and appearing objects in the visual scene. The object classes and the ordering of their disappearance induce possible assembly structures of complex objects that are placed back into the scene. These possible structures are modeled by *planning rules*. For example, the system detects in three sequential steps the disappearance of a bolt, a bar, and a cube. Now, if the appearance of a new object is detected a first top down hypothesis of a complex object can be generated consisting of the bolt, the bar that is plugged onto it, and the cube.

Both partial recognition results are fused into one final structure of complex objects in the scene.

This structural information is utilized by the integration component that considers the verbal context in two different ways. First, elementary objects that are structurally described in an utterance can be identified, e.g. *"the green cube that is connected to the long bar"*. The resulting task is a variant of Task 2:

Task 4 (*mp-struct*) *Which objects are most probably denoted by a verbal structural description?*

Verbal descriptions may also help to find unrecognized assembly relations in the scene.

Secondly, detected complex objects are good candidates for linking them with unknown object nouns, e.g. *"Please give me the part in front of the plane's tail."*. Here, the reference object can be probably identified as a complex object which constrains the identification of the intended one. Additionally, the found linkage of the complex object and the unknown object noun can be recorded as an implicit naming that can be used in the interpretation of the further construction process. This is a second variant of Task 2:

Task 5 (*mp-name*) *Which objects are most probably named by object nouns with unknown semantics?*

4.3.3 Speech understanding and vision results

An important aspect for the integration of speech processing and vision results is the reliability of them. On the speech side, the quality of the recognition component can be measured by the ***word accuracy*** (WA) [Lee89]:

$$WA = \frac{\text{\#words} - (\text{\#substitutions} + \text{\#insertions} + \text{\#deletions})}{\text{\#words}} \cdot 100\% \qquad (4.1)$$

The recognized word sequence is checked against a transcribed reference sequence of words. Word substitutions, insertions, and deletions are counted and normalized by the total number of words of the reference sequences. If the word accuracy is 100% no speech recognition errors occurred. If many insertion errors occur the word accuracy can even

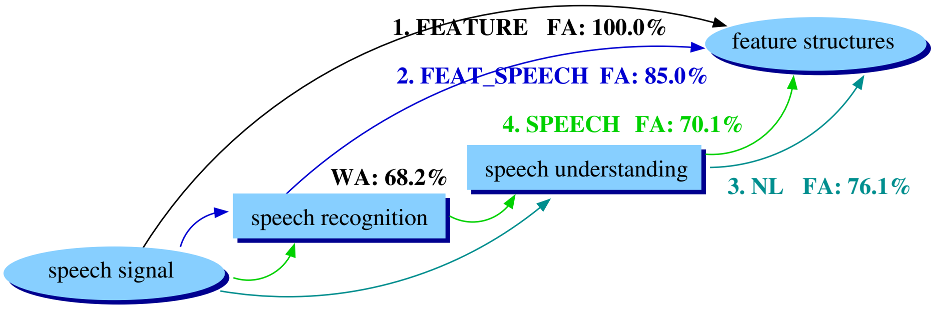


Figure 4.8: Word and feature accuracies for the *Select-Obj* set.

be negative. The word accuracy neither takes account of errors that happen on the level of speech understanding, like unmodeled sentence structures, nor considers whether the recognized word is relevant for further processing. In order to capture the influence of speech recognition *and* understanding on the interpretation of verbal object descriptions, a similar measure can be defined that counts features instead of words, i.e. the ***feature accuracy*** (FA):

$$FA = \frac{\#\text{features} - (\#\text{substitutions} + \#\text{insertions} + \#\text{deletions})}{\#\text{features}} \cdot 100\% \qquad (4.2)$$

For each object description that was detected in the utterances the features used to query the intended object are counted and compared to a reference transcription. In contrast to the word accuracy, the feature accuracy is invariant to the ordering of features. However, a single misrecognized word may lead to a misinterpreted sentence structure that can result in a loss of several features. Such effects can be measured by calculating the accuracy on an object level. An object description is correctly recognized if all its features are correct. If no reference object was detected this is counted as an object deletion. The ***object accuracy*** is defined similar to the word and feature accuracies.

The speech recognizer used a bigram language model that has been estimated from the transcribed construction dialogs of experiment 3 (Human-Machine) and a hand modeled partial grammar. The recognition lexicon contained 1193 words.

Fig. 4.8 visualizes the different processing ways that have been evaluated. In ***FEATURE*** the utterances have been transcribed into feature structures. This is the reference data set. In ***FEAT_SPEECH*** the *recognized* word sequences have been transcribed into feature structures. These two sets assume a perfect understanding component. The next two sets ***NL*** and ***SPEECH*** measure the performance of the understanding component with regard to verbal object descriptions. In ***NL*** the transcribed text and in ***SPEECH*** the speech signals are processed. Table 4.1 presents the recognition[5] and understanding results of the system for the two evaluation sets:

- *Select-Obj* set:

[5] The parameters of the speech recognizer had been optimized for the two different sets. The *Select-Obj* results were produced by combining acoustic, bigram, and grammar scores. For the *Select-Rel* results only acoustic and grammar scores were combined.

Experim. 5 (Select-Obj)	#utt	total	sub	del	ins	corr	acc
words	453	2394	**22.9**	4.5	4.5	72.66	**68.2**
features (FEAT-SPEECH)	447	969	**4.0**	4.8	6.2	91.2	**85.0**
objects (FEAT-SPEECH)	447	475	**20.4**	0.6	1.7	79.0	77.3
features (NL)	397	856	0.9	**21.6**	1.4	77.5	**76.1**
features (SPEECH)	373	822	3.2	**23.2**	3.5	73.2	**70.1**

Experim. 6 (Select-Rel)	#utt	total	sub	del	ins	corr	acc
words	174	1396	**15.0**	3.9	1.7	81.1	**79.5**
features (FEAT-SPEECH)	141	508	**1.6**	6.5	4.7	91.9	**87.2**
objects (FEAT-SPEECH)	141	283	10.3	**4.6**	0.0	85.2	85.2
features (NL)	173	639	0.0	6.7	0.0	93.3	**93.3**
features (SPEECH)	141	508	3.2	**21.5**	3.7	75.4	**71.7**
objects (SPEECH)	141	283	15.2	**14.1**	0.0	70.7	70.7

Table 4.1: Speech recognition results on the evaluation sets using a recognition lexicon of 1193 words. The *Select-Obj* set includes utterances from 10 different speakers, the *Select-Rel* set those from 6 different speakers. The ’total’ column gives number of words, features, or object descriptions included in each evaluation set.

1. Although the substitution rate on word level is very high (22.9%) the recognition of relevant features is nearly stable (only 4.0% substitutions).
2. Nevertheless, 20.4% of the verbal object descriptions are affected by speech recognition errors.
3. In 6 out of 453 utterances either no relevant feature for an object description was detected or the intended object is interpreted as the reference object because a spatial relation was inserted.
4. More than 20% of the features were not detected by the understanding component due to unmodeled sentence structures.

- *Select-Rel* set:

  1. Again, the sustitution rate on word level is much higher (15.0%) than that on the feature level (1.6%).
  2. This time the word accuracy (79.5%) is significantly higher than that of the previous test set. The reason is that most of the occuring sentence structures are covered by the partial grammar that is integrated in the speech recognizer.
  3. As a consequence the feature accuracy of the NL set is very high (93.3%).
  4. The speech recognition errors break down the performance of the understanding component (FA 71.7%). The reason is a loss of detected reference objects (object deletions 14.1%).

The loss of features and reference objects that were noted for the understanding component will significantly affect the integration of speech and images. Thus, any object

(types)	#objects	false	not-detected	inserted	correct	DA
Experim. 5 (Select-Obj)	165	**17.6**	1.8	6.1	80.6	**74.6**
Experim. 6 (Select-Rel)	86	**15.1**	2.3	11.6	82.6	**70.9**
(colors)	#objects	false	not-detected	inserted	correct	DA
Experim. 5 (Select-Obj)	165	4.9	1.8	6.1	93.3	**87.3**
Experim. 6 (Select-Rel)	86	9.3	2.3	11.6	88.4	**76.7**
(color+type)	#objects	false	not-detected	inserted	correct	DA
Experim. 5 (Select-Obj)	165	17.6	1.8	6.1	80.6	**74.6**
Experim. 6 (Select-Rel)	86	22.1	2.3	11.6	75.6	**64.0**

Table 4.2: Object recognition results of elementary objects on the evaluation sets. The recognizer has to distinguish 11 different types and 10 different colors. If both are correctly classified the object is counted 'correct' in the third tabular. The *Select-Obj* set consisted of 11 different images, the *Select-Rel* set consisted of 6 different scenes.

identification result will mainly reflect the performance of the understanding component, not the integration framework. Therefore, the ***FEATURE*** and ***FEAT_SPEECH*** sets will be used in order to generate the quantitative evaluation results of this thesis (Chap. 6).

The scheme of considering substitutions, insertions, and deletions can also be applied to the visual object recognition results. Insertions and deletions capture segmentation errors. Substitutions count wrong color or type classifications of objects. Thus, the ***detection accuracy*** is defined as:

$$DA = \frac{\#\text{objects} - (\#\text{false-classification} + \#\text{inserted-object} + \#\text{not-detected})}{\#\text{objects}} \cdot 100\% \tag{4.3}$$

The recognition results of elementary objects are presented in Table 4.2. The error rates reveal the same tendencies in both evaluation sets:

1. In both sets nearly all objects were detected.

2. Most of the inserted objects are small *unknown* object regions that are generated due to a over-segmentation of objects.

3. Most errors occure during the classification of types (17.6%, 15.1%).

4. Consequently, the detection accuracy of colors (87.3%, 76.3%) is higher that of types (74.6%, 70.9%).

The recognition accuracy of complex objects is more difficult to measure because structural descriptions must be compared and the accuracy strongly depends on the complexity of the assembled object. Qualitatively, complex objects are correctly detected if all elementary object parts have been recognized correctly.

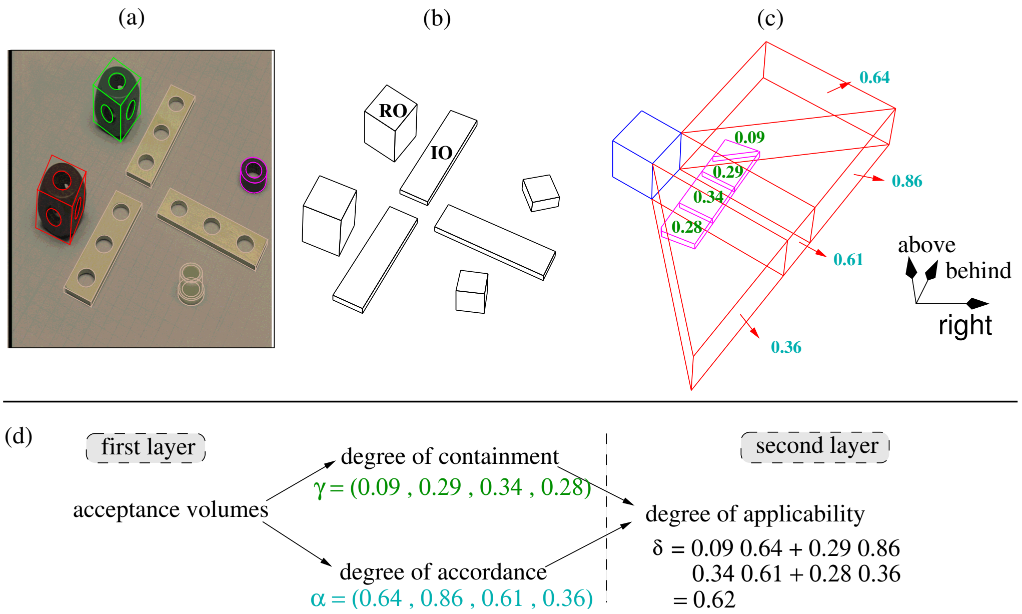


Figure 4.9: Computation of 3-d projective relations. (a) shows adapted CAD-models of the recognized objects. In (b) they are approximated by bounding boxes. (c) shows the representation of the first layer with regard to the projective relation *'right'* between the objects *IO* and *RO*. (d) The degree of applicability δ is calculated on the second layer.

4.4 Spatial Modeling

The spatial model used in the speech and vision integrating component is based on the 3-d spatial model proposed by Fuhr et. al [FSSS97] which will be outlined in the next subsection. In this thesis it will be modified and extended in order to capture more relevant aspects of spatial inference in the baufix®domain. The proposed spatial model is not intended to be completely general but it preserves the flexibility of the domain and is formulated on a mathematically sound geometrical basis.

4.4.1 A model for 3-d projective relations

The 3-d spatial model proposed by Fuhr et. al [FSSS97] combines a space partitioning approach with a fuzzy scoring scheme and therefore combines neat and scruffy aspects (cf. Sec. 2.5.2). It models the six projective relations *left, right, in-front-of, behind, above*, and *below*. A projective relation is a directed relation between a reference object (*RO*) and an intended object (*IO*) with regard to a reference frame (*RF*). The spatial model calculates the ***degree of applicability*** of a projective relation from the geometric properties of the intended and reference object. The influence of the reference frame is considered in a two-layered process.

First, The object shapes are approximated by bounding boxes (Fig. 4.9b). The bounding box of the reference object is used in order to partition the three-dimensional space into a constant number n of infinite ***acceptance volumes***[6] each of which has an associated direction vector (Fig. 4.9c). The relevance of each acceptance volume $AV_i^{RO} = \{\mathcal{V}_i^{RO}, \vec{d}_i^{RO}\}$, $i = 1 \ldots n$ with regard to an intended object that occupies the volume $\mathcal{V}^{IO}$ is defined by the ***degree of containment*** γ:

$$\gamma(\mathcal{V}_i^{RO}, \mathcal{V}^{IO}) = \frac{|\mathcal{V}_i^{RO} \cap \mathcal{V}^{IO}|}{|\mathcal{V}^{IO}|} \tag{4.4}$$

Up to this stage, the calculation is independent of a selection of the reference frame.

In parallel, a second relevance value is calculated for each acceptance volume that represents the influence of the reference frame. The reference frame determines the three-dimensional direction $\vec{d}^{RF}$ that is denoted by a projective relation, like *'right'*. The ***degree of accordance*** α defines the relevance of each acceptance volume $AV_i^{RO} = \{\mathcal{V}_i^{RO}, \vec{d}_i^{RO}\}$ with regard to this direction:

$$\alpha(\vec{d}_i^{RO}, \vec{d}^{RF}) = \begin{cases} 1 - \frac{2}{\pi} \cdot \arccos(\langle \vec{d}_i^{RO}, \vec{d}^{RF} \rangle) & \text{, if } \langle \vec{d}_i^{RO}, \vec{d}^{RF} \rangle > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{4.5}$$

These two measurements constitute the first layer of the spatial model.

The second layer combines these two fuzzy measurements. The relevance relationships of the acceptance volumes are combined to the ***degree of applicability*** δ of a projective relation p between the intended object (IO) and the reference object (RO):

$$\delta(p, IO, RO, RF) = \sum_{AV_{i=1\ldots n}^{RO} = \{\mathcal{V}_i^{RO}, \vec{d}_i^{RO}\}} \gamma(\mathcal{V}_i^{RO}, \mathcal{V}^{IO}) \cdot \alpha(\vec{d}_i^{RO}, \vec{d}^{RF}) \tag{4.6}$$

where $AV_{1\ldots n}^{RO}$ is the space partitioning with regard to the reference object,
$\vec{d}^{RF}$ is the three-dimensional direction of the projective relation p
with regard to the reference frame RF.

The degree of containment γ can be computed before the reference frame and the named projective relation are known. It can even be re-used for the calculation of different projective relations. Additionally, the same spatial model can be used to specify the spatial area where an intended object is expected by thresholding the degree of accordance and joining all acceptance volumes that remain relevant. This aspect may be exploited by an expectation-driven vision strategy.

4.4.2 The spatial model in two dimensions

The 3-d spatial model described above has some drawbacks that motivate to switch to a simpler model that is defined in only two dimensions:

- In order to apply the 3-d model to an object pair, the 3-d shape of these objects must be fully reconstructed in order to compute the bounding boxes. Unconstrained or

[6] A typical selection for n is 79. One for the bounding box itself, one for each side of it, two for each edge, and six for each corner.

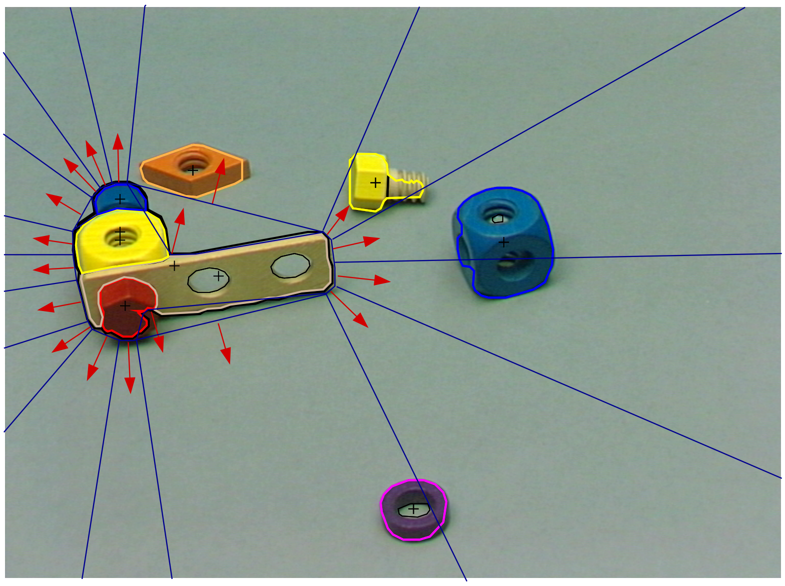


Figure 4.10: Space partitioning in the 2-d model.

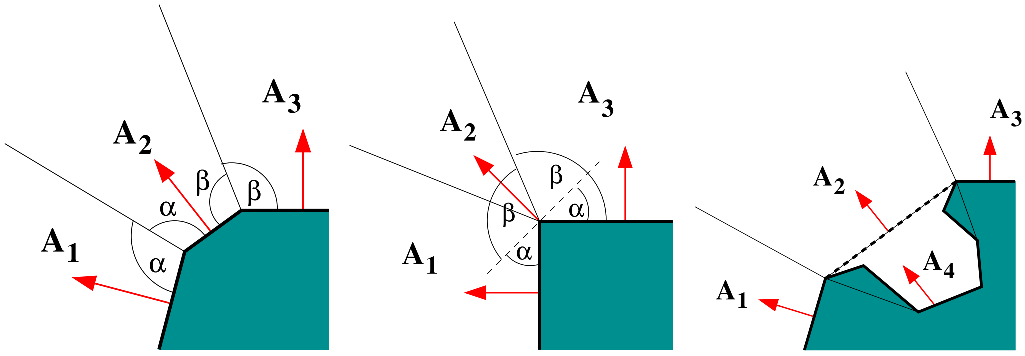
only coarsely constrained rotational degrees of freedom introduce inaccuracies to the calculations in the model. Objects in the scene that have only been detected on a blob level are difficult to include in the spatial model.

- The bounding box is a too coarse shape abstraction especially for complex objects. Complex objects consist of a set of elementary baufix®objects that have been joined by screwing or plugging. They have much more complicated shapes that need not be convex.
- The 3-d model does not consider the plane of the table as an influencing factor. Most uses of projective relations are applied to two objects that both lie on the same table plane. Therefore, these cases can be adequately handled in two dimensions.

The solution proposed in this thesis contributes to these points by calculating 2-d relations on the image plane and relaxing the shape descriptions of the intended and reference object to unconstrained polygons. Bounding boxes are substituted by the outline polygons, acceptance volumes by acceptance areas, and associated 3-d direction vectors by 2-d direction vectors. Instead of a partitioning of the 3-d space, a partitioning of the image plane is calculated.

Definition of acceptance areas

Fig. 4.10 shows the space partitioning of the 2-d spatial model. All objects in the scene are represented by the outline polygons of the corresponding object regions in the image. These have been calculated by the object recognition components of the system. In



(a) definition of the acceptance area

(b) acceptance areas at wide angles

(c) acceptance areas for a concave outline

Figure 4.11: Definition of acceptance areas for the 2-d spatial model.

Fig. 4.10 the complex object is assumed to be the reference object of a projective relation. The space partitioning is based on the convex hull of the object outline:

- For each edge of the convex hull one acceptance area is defined that is bounded by this edge and two lines that bisect the outer angles $2\alpha, 2\beta$ to the adjacent edges of the polygon (see Fig. 4.11(a)). The acceptance area is attached with a vector that points away from the object in a direction that is defined orthogonal to the polygon edge.

- If the outer angle is greater than a specified threshold an additional acceptance area is inserted with an attached direction vector that bisects the angle (Fig. 4.11(b)). Consequently, the adjacent acceptance areas are defined with regard to a reduced outer angle 2β.

- For each concave section of the outline an acceptance area $\mathcal{A}_4$ is defined that inherits the direction vector of the adjacent *'convex'* acceptance area $\mathcal{A}_2$ (Fig. 4.11(c)). The area between the convex hull and the object outline is approximated by a convex polygon.

The representation layers of the 2-d spatial model

Based on the space partitioning the representation layers are defined similarly to the 3-d model. The ***degree of containment*** measures the relative area of the intended object *IO* that is contained by an acceptance area $\mathcal{A}_i^{RO}$:

$$\gamma(\mathcal{A}_i^{RO},\mathcal{A}^{IO}) = \frac{|\mathcal{A}_i^{RO} \cap \mathcal{A}^{IO}|}{|\mathcal{A}^{IO}|}, \tag{4.7}$$

where $\mathcal{A}^{IO}$ is the image area occupied by the intended object,
$\mathcal{A}_i^{RO}$ is the i-th acceptance area of the reference object.

In order to calculate the ***degree of accordance***, the reference frame is projected onto the image plane. Subsequently, the angle differences between the direction of the projected spatial relation and the 2-d direction vector of the acceptance areas are computed:

$$\alpha(\vec{d}_i^{RO},\vec{d}^{RF}) = \begin{cases} 1-2\cdot\frac{\arccos(\langle\vec{d}_i^{RO},\vec{d}_{2D}^{RF}\rangle)}{\pi} & ,\text{if } \langle\vec{d}_i^{RO},\vec{d}_{2D}^{RF}\rangle > 0 \\ 0 & ,\text{otherwise} \end{cases} \tag{4.8}$$

where $\vec{d}_{2D}^{RF} = Proj[\underline{R},\underline{t}](\vec{d}^{RF})$,
$\vec{d}_i^{RO}$ is the 2-d direction vector of the acceptance area,
$\vec{d}^{RF}$ is the specified 3-d direction in world coordinates with regard to the reference frame RF,
$Proj$ projects a 3-d direction vector defined in world coordinates onto the image plane,
$\underline{R}$ is the rotation matrix from world to camera coordinates,
$\underline{t}$ is the translation vector from world to camera coordinates.

Fig. 4.12 shows the first representation layer of the 2-d spatial model for an object pair and the projective relation *'behind'*. The numbers in the acceptance areas denote the degree of accordance, numbers placed in the area of the intended object denote the degree of containment.

On the second layer the ***degree of applicability*** is calculated similarly to the 3-d model (Eq. 4.6). For the example in Fig. 4.12 it yields:

$$\delta(behind, IO, RO, RF) = 0.6\cdot 0.7 + 0.15\cdot 0.7 = 0.525$$

The spatial model for projective relation can even be used for processing combinations of the primitive relations *left, right, in-front-of, behind, above, below*. In this case the denoted 3-d direction $\vec{d}^{RF}$ is a linear combination of the base vectors of the reference frame.

The trinary spatial relation *'between'*

The spatial model for projective relations can be used in order to calculate the degree of applicability of the trinary relation *'between'* that takes one intended object (IO) and two reference objects (RO_1, RO_2) as arguments. The relation $between(IO, RO_1, RO_2)$ is defined considering the following principles:

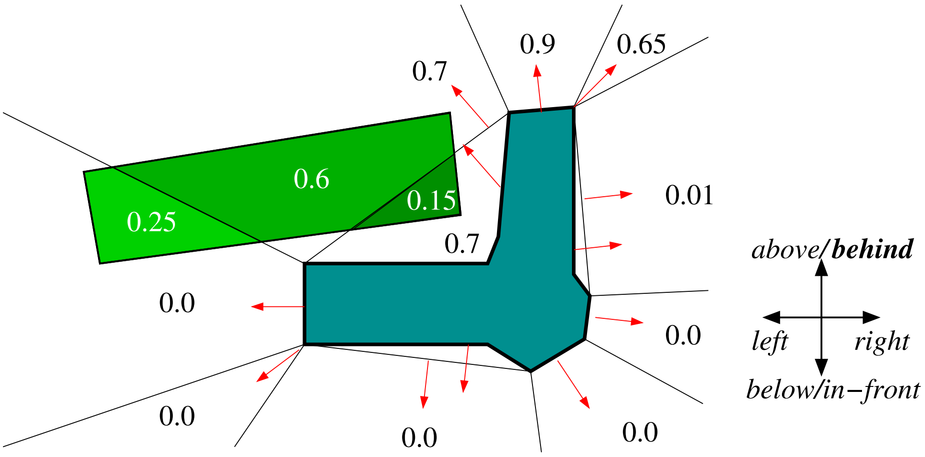


Figure 4.12: First-level representation of the 2-d spatial model for *'behind'*.

- It is independent of the reference frame.
- It is symmetric with regard to an exchange of the reference objects.
- The two reference objects of a relation *between* are located in opposite directions.

The degree of applicability is calculated in two steps. First, the direction between each reference object and the intended object is computed separately by integrating the acceptance volumes that are covered by the area of the intended object:

$$\vec{d}^{IO,RO_i} = \sum_{j=1\ldots n} \gamma(\mathcal{A}_j^{RO_i}, \mathcal{A}^{IO})\, \vec{d}_j^{RO_i}, \quad i = 1,2 \tag{4.9}$$

where $\{\{\mathcal{A}_j^{RO_i}, \vec{d}_j^{RO_i}\} | j = 1\ldots n\}$ is the space partitioning of reference object RO_i, $\mathcal{A}^{IO}$ is the object area of the intended object.

In a second step the degree of accordance α (Eq. 4.5) between the vector $\vec{d}^{IO,RO_1}$ and the inverted vector $-\vec{d}^{IO,RO_2}$ defines the degree of applicability:

$$\delta(between, IO, RO_1, RO_2) = \alpha\left(\frac{\vec{d}^{IO,RO_1}}{||\vec{d}^{IO,RO_1}||}, -\frac{\vec{d}^{IO,RO_2}}{||\vec{d}^{IO,RO_2}||}\right) \tag{4.10}$$

4.4.3 The neighborhood graph

So far the projective spatial relations have been defined without considering any context objects. Consequently, the applicability of a specified spatial relation is very loosely constrained in table scenes with many objects. In Fig. 4.13 the system interpretation of the utterance *"Please take the object to the left of the long green bolt"* would be an arbitrary selection of one of ten possible objects. However, a short introspection yields a single object, the seven-holed bar, as the object that was denoted by the speaker.

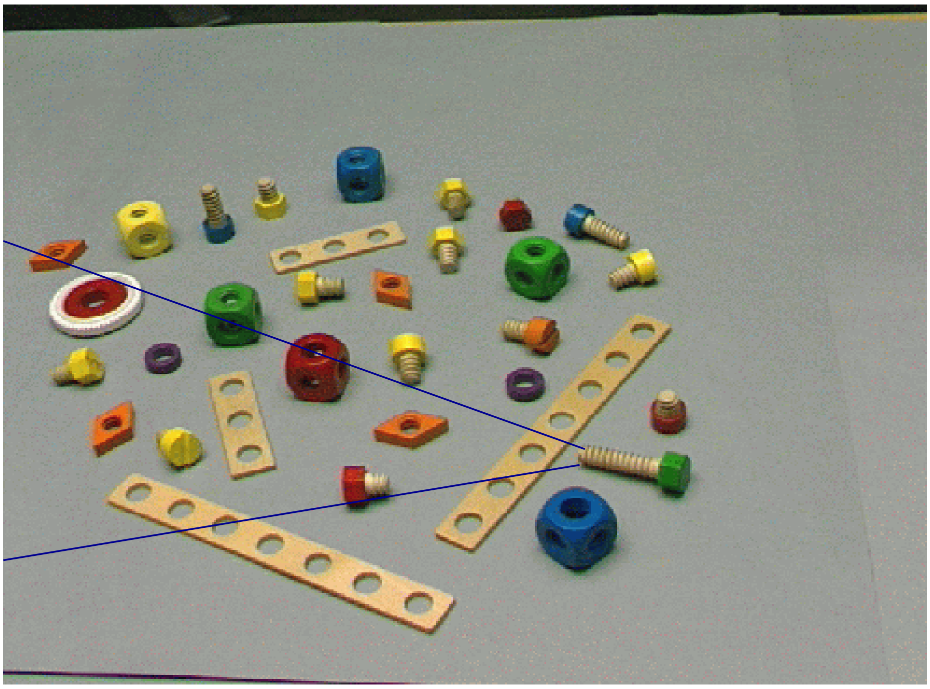


Figure 4.13: Projection of the spatial relation *’left’* with regard to the reference object *long green bolt*.

The assumption consequently introduced to the spatial model is the concept of ***neighborhood*** [SWBPK99, WBPS$^+$99, WBPK$^+$99]:

> *The speaker will select a reference object in the neighborhood of the intended object when a spatial relation is specified.*

In Fig. 4.13 the seven-holed bar is neighboring the *long green bolt*. The other possible objects do not neighbor the bolt because the bar is placed in between.

The seven-holed bar ***separates*** these objects from the *long green bolt*.

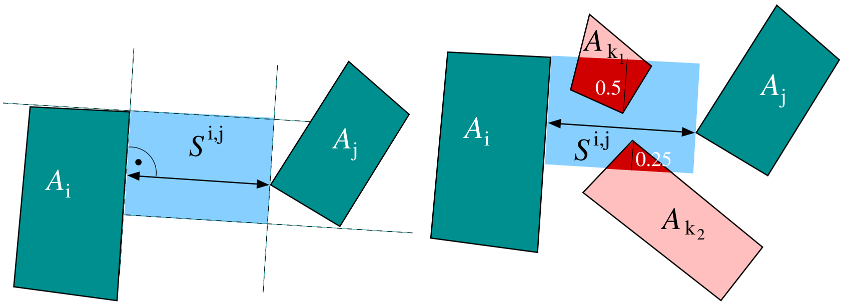
In this sense, the neighborhood definition is based on ***separation***:

Def. 6 (neighborhood) *Two objects are in neighborhood if there is no object in between that separates them.*

The ***separation*** predicate is calculated using a geometrical concept that is defined on the image plane and can therefore be applied to any object that is represented on a blob level:

Def. 7 (separation) *Let $\mathcal{A}_1,\ldots\mathcal{A}_n$ be the image areas occupied by the scene objects $O_1,\ldots,O_n$. The separation predicate between objects O_i,O_j is defined by thresholding the **degree of separation** Sep between two object areas:*

$$Sep(\mathcal{A}_i,\mathcal{A}_j,\{\mathcal{A}_k|k=1\ldots n,k\neq i,k\neq j\}) > \Theta_{Sep} \tag{4.11}$$



(a) Area in between

(b) Degree of separation

Figure 4.14: The degree of separation of an object pair is determined on the *area in between* (a). The first extension of this rectangular area is given by the shortest line between the two outlines of the object areas. The second orthogonal extension is defined by the minimum of the two object extensions in this direction. In (b) the object areas $\mathcal{A}_i$ and $\mathcal{A}_j$ are separated by two context objects. The separation degree introduced by object area $\mathcal{A}_{k_1}$ is 0.5. That of object area $\mathcal{A}_{k_2}$ is 0.25.

The degree of separation is determined on a rectangular area $\mathcal{S}^{i,j}$ in between the object areas $\mathcal{A}_i, \mathcal{A}_j$ (Fig. 4.14(a)):

$$Sep(\mathcal{A}_i, \mathcal{A}_j, \{\mathcal{A}_k | k = 1 \ldots n, k \neq i, k \neq j\}) = \frac{\| \bigcup_{k=1\ldots n, k\neq i, k\neq j} \mathcal{A}_k \cap \mathcal{S}^{i,j} \|_{i,j}}{\|\mathcal{S}^{i,j}\|_{i,j}} \tag{4.12}$$

where $\|.\|_{i,j}$ measures the maximal extension of an area in an orthogonal direction to the shortest line between the outlines of the areas $\mathcal{A}_i, \mathcal{A}_j$ (Fig. 4.14(b)).

There are some cases when the rectangular area $\mathcal{S}^{i,j}$ between two objects degenerates. If one of the two objects O_1, O_2 has a very small size and the distance between these objects is very large, there is only a very thin corridor considered for calculation of the degree of separation. Therefore, the ratio between the width and the length of such a corridor is limited by the model. If the ratio is smaller than the limit the width of the area $\mathcal{S}^{i,j}$ is enlarged. Thus, the distance between two objects is considered indirectly.

The separation criterion introduces a ***neighborhood graph*** to the scene representation. The nodes of this graph are detected object regions, and the edges are neighborhood relations. Fig. 4.15 presents this graph for an image example. The applicability of projective spatial relations is constrained by the neighborhood graph. How the direction and separation criteria are translated into a probability measure will be discussed later in section 4.5.4.

4.4.4 Localization attributes

Another kind of local information is introduced by specifying the relative position in the scene without explicitly mentioning a reference object:

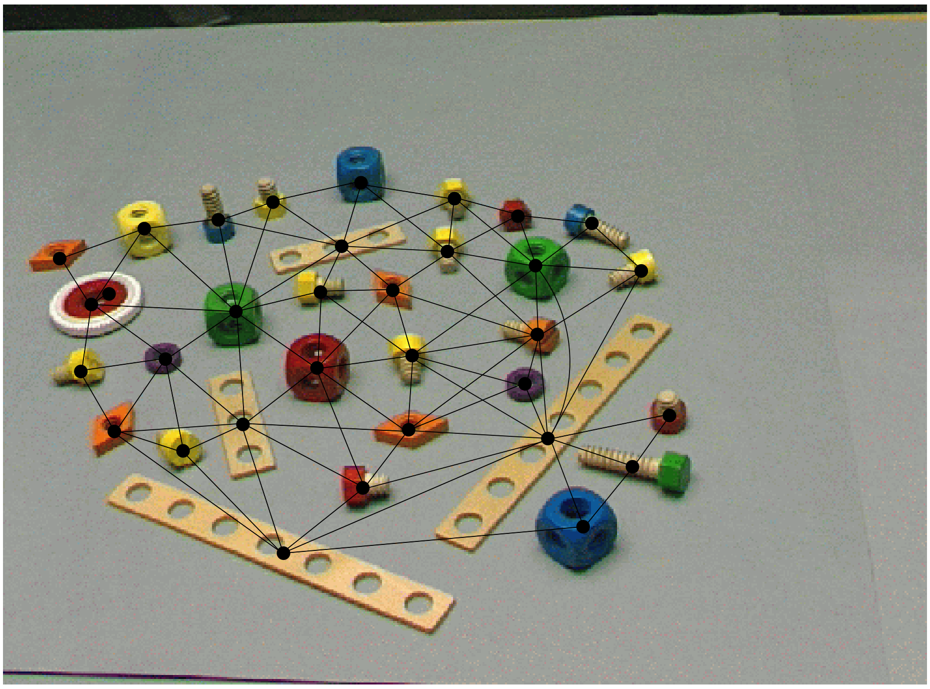


Figure 4.15: The neighborhood graph for an example image.

”... the short yellow bolt in the middle.”
”... the blue cube in the front.”

Such spatial descriptions are related to the positions of all objects in the scene. Therefore, it is very difficult to specify an adequate space partitioning.

In order to capture the meaning of such local descriptions, a simple potential field model is constructed. The potential field is normalized by the positions of the most outward objects. The potential field is oriented in the direction of the specified attributes, e.g. *loc* = *on-the-right* (Fig. 4.16(a)):

$$\delta(loc, \mathcal{A}_i, RF, \{\mathcal{A}_k | k = 1 \ldots n\}) = c \cdot tanh(a \cdot x) + 0.5 \tag{4.13}$$

where $x = \dfrac{\langle \vec{m}_i | \vec{d}_{2D}^{RF} \rangle - (\max_k \langle \vec{m}_k | \vec{d}_{2D}^{RF} \rangle + \min_k \langle \vec{m}_k | \vec{d}_{2D}^{RF} \rangle)/2}{\max_k \langle \vec{m}_k | \vec{d}_{2D}^{RF} \rangle - \min_k \langle \vec{m}_k | \vec{d}_{2D}^{RF} \rangle}$,

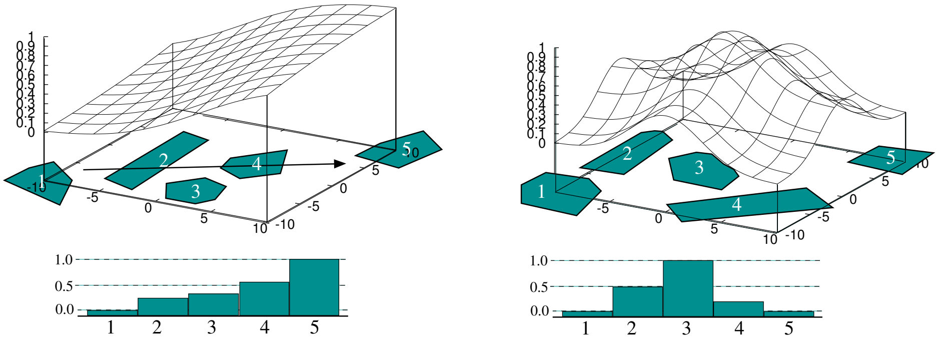
$\vec{m}_i$ is the center of object area $\mathcal{A}_i$,

$\vec{d}_{2D}^{RF}$ is the specified direction *loc* with regard to the reference frame *RF* that was projected onto the image plane,

a defines the gradient of the potential function,

c is a normalizing constant that scales the function to the interval $[-0.5; 0.5]$ between $\min_k \langle \vec{m}_k | \vec{d}_{2D}^{RF} \rangle$ and $\max_k \langle \vec{m}_k | \vec{d}_{2D}^{RF} \rangle$.

If a localization attribute that defines a direction, e.g. *’on-the-right’*, is combined with the attribute *’in-the-middle’*, the potential function is multiplied by a Gaussian function that



(a) ”On the right”

(b) ”In the middle”

Figure 4.16: Potential fields of localization attributes. The numbered regions are detected objects in the scene. The specified direction vector depends on the reference frame.

is oriented orthogonal to the direction of the other attributes, e.g.:

$$\delta(\textit{on-the-middle-right}, \mathcal{A}_i, RF, \{\mathcal{A}_k | k = 1 \dots n\}) \\ = \delta(\textit{on-the-right}, \mathcal{A}_i, RF, \{\mathcal{A}_k | k = 1 \dots n\}) \cdot \exp(-a \cdot y^2) \qquad (4.14)$$

where y is defined similarly to x in Eq. 4.13 with regard to the orthogonal direction $\vec{d}_{2D}^{RF\perp}$,
a defines the gradient of the Gaussian function.

The potential field of the isolated use of *’in-the-middle’* is shown in Fig. 4.16(b). Here, two Gaussian functions that are oriented with the *left-right* and *behind-infront* directions of the reference frame are summed up:

$$\delta(\textit{in-the-middle}, \mathcal{A}_i, RF, \{\mathcal{A}_k | k = 1 \dots n\}) = \frac{1}{2}[\exp(-a \cdot x^2) + \exp(-a \cdot y^2)] \qquad (4.15)$$

where x, y are defined similar to Eq. 4.13, 4.14,
a defines the gradient of the Gaussian functions.

4.4.5 Summary

The spatial model proposed in the previous subsections captures most of the relevant aspects which have been discussed in Sec. 2.5.2.

- **Dimensionality:** When the position of an intended object is verbally specified the description mostly refers to the 2-d plane of the table. Therefore, locations and spatial relations can be interpreted in two dimensions. The advantage of such an approach is that the recognition of the objects in the scene is no pre-requisite. The shape properties of the polygon outline of detected object regions is sufficient for the application of a 2-d model. The disadvantage of this approach is that specified directions may be ambiguous and that perspective occlusions may introduce erroneous interpretations.

- **Topology:** The most relevant contact relations in the baufix®-domain are ***mounting relations*** that are captured by the assembly model of ***complex objects*** (see Sec. 4.3.2).

- **Position and orientation:** The relative pose between two objects is described by a discrete space partitioning approach with a fuzzy measurement.

- **Scale:** In the baufix®domain there are two different levels of scale: separate objects that constitute the table scene and assembled parts that constitute the structure of a complex object. The proposed spatial model is designed for the description of table scenes. It makes the assumption that – instead of distance – the separation of objects by other context objects is the most important influence on the applicability of spatial relations. The second level is dominated by mounting relations that are captured by the syntactical approach for the recognition of complex objects.

- **Shape:** The shape of the intended object and reference object is approximated by the outline polygon of the corresponding object area in an image. The proposed spatial model even considers polygons with concave shapes. The shape of complex objects can thereby be represented in an adequate way. The outline polygon of the reference object determines the space partitioning. The shape of the intended object is considered in the calculation of the degree of containment.

- **Multiple objects:** The spatial model considers the configuration of context objects by introducing the neighborhood concept. The reference object will be selected in the neighborhood of the intended object. Thus, if other objects are placed between them the usage of a spatial relation between them is less probable.

- **Time:** Time is not considered because the spatial model should only describe static scenes.

- **Reference frame:** The application of the spatial model is designed as a two layered process. The first layer is independent of the reference frame whereas the second layer considers the influence of it. The meaning of a spatial relation can thereby be calculated very easily with regard to different reference frames. The selection of the appropriate reference frame is external to the model.

The spatial model proposed so far has some aspects that are domain-specific, like the definition of the neighborhood concept that is based on the separation of objects. Other aspects like the shape of involved objects and the space partitioning are more general. The previous work of Fuhr et al. shows that the same concepts can be applied even in three dimensions.

4.5 Object Identification using Bayesian Networks

The identification of those scene objects that are denoted in a spoken instruction is the central task in the multi-modal interpretation process of the system [WBPK+99, WBPS+99]. It comprises the solution of the ***correspondence problem*** (see 2.5) and has been introduced as the ***mp-objs*** task (Task 2).

The input data is provided by the speech understanding and vision components. The visual scene representation *V_SCENE* consists of the set of detected objects:

V_SCENE:	*V_OBJS*	$\subset \{so \mid so \in V\_OBJECT\}$
V_OBJECT:	*OBJ_POLY*	2-d object outline in the image
	OBJ_CLASS	$\in \{$ *3/5/7-holed-bar, cube, rim, rhomb-nut, tire, socket, thin/thick washer, bolt, assembly, undef* $\}$
	COL_CLASS	$\in \{$ *red, yellow, orange, blue, green, violett, wooden, ivory-colored, white, black, undef* $\}$
	SUB_PARTS	$\subset \{sp \mid sp \in V\_OBJECT \cup V\_OBJ\_PART\}$

V_OBJ_PART is defined similarly to *V_OBJECT* but the set of possible object classes is different, e.g. {*head, thread, bar-body, bar-hole*, ...}. These parts have no other subparts.

The verbal representation of a spoken utterance *S_UTTERANCE* consists of a number of object descriptions and a set of relations that are defined between them:

S_UTTERANCE:	*S_OBJS*	$\subset \{uo \mid uo \in S\_OBJECT\}$
	S_RELS	$\subset \{ R_n(o_1,\ldots,o_n) \mid o_i \in S\_OBJS, R_n \in \{Left_2, Right_2, In\text{-}Front\text{-}Of_2, Behind_2, Above_2, Below_2, Between_3, Connected_2, Part\text{-}Of_2\}$
S_OBJECT:	*OBJ_NOUN*	$\in \{$ *bar, three/five/seven-holed bar, cube, rim, bolt, rim, nut, part, object, tail, plane, ...* $\}$
	ATTRIBUTES	$\subset \{$ *red, blue, dark, thin, large, round, long, ...* $\}$
	LOC_ATTRS	$\subset \{$ *on-the-left, on-the-right, in-the-front, in-the-back, in-the-middle,* $\}$

In the following, these two representations will be related by a Bayesian network.

Before starting with the detailed modeling proposed in this thesis some aspects of the previous work of Gudrun Socher [Soc97, SSP00] will be presented that provides the basis for my own work.

4.5.1 Previous work

Socher et al. divide the identification process into two separate steps. First, each object that is mentioned in an utterance is identified separately. Secondly, mentioned spatial relations are checked in order to restrict the set of selected objects.

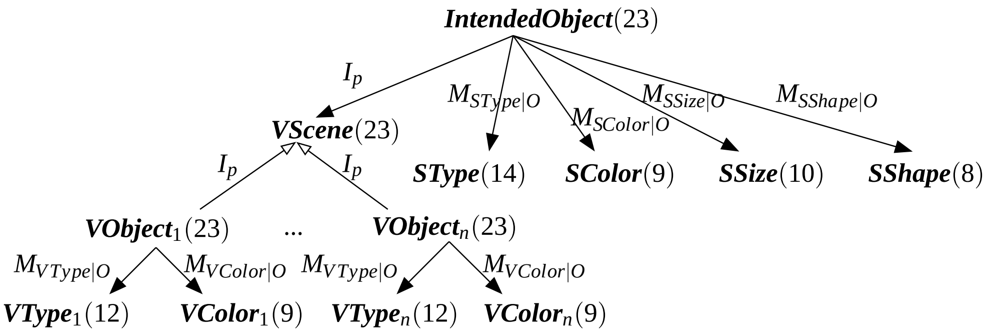


Figure 4.17: Structure of the Bayesian network proposed by Socher et al. [SSP00].

The first step is realized using the Bayesian network presented in Fig. 4.17. The leaves of the tree-structured net are the evidential nodes: ***SType, SColor, SSize, SShape*** for evidences extracted from speech and $\boldsymbol{VType}_i$, $\boldsymbol{VColor}_i$, $i = 1 \ldots n$, for visual evidences. Nouns that were frequently used in experiment 5 (Select-Obj) define the possible labels of the variable ***SType***, frequently applied color adjectives are collected in the variable ***SColor***[7]:

$$\begin{aligned}\boldsymbol{SType} \in \{&\text{Objekt}, \text{Leiste}, \text{Schraube}, \text{Dreilochleiste}, \text{Fünflochleiste}, \text{Siebenlochleiste}, \text{Würfel},\\ &\text{Raute}, \text{Felge}, \text{Reifen}, \text{Rundkopfschraube}, \text{Sechskantschraube}, \text{Unterlegscheibe}\}.\\ \boldsymbol{SColor} \in \{&\text{weiß}, \text{rot}, \text{gelb}, \text{orange}, \text{blau}, \text{grün}, \text{violett}, \text{holz}, \text{elfenbein}\}\end{aligned}$$

The shape and size adjectives used in experiment 4 (WWW) define the labels of the variables ***SSize*** and ***SShape***[7]:

$$\begin{aligned}\boldsymbol{SSize} &\in \{\text{klein}, \text{groß}, \text{kurz}, \text{lang}, \text{mittellang}, \text{mittelgroß}, \text{dick}, \text{dünn}, \text{schmal}, \text{hoch}\}\\ \boldsymbol{SShape} &\in \{\text{rund}, \text{eckig}, \text{länglich}, \text{sechseckig}, \text{viereckig}, \text{rautenförmig}, \text{flach}, \text{rechteckig}\}\end{aligned}$$

The possible values of the variables $\boldsymbol{VType}_i$ are the twelve object classes that are distinguished by the recognizer for elementary objects:

$$\begin{aligned}\boldsymbol{VType}_i \in \{&\mathit{3\text{-}holed\text{-}bar}, \mathit{5\text{-}holed\text{-}bar}, \mathit{7\text{-}holed\text{-}bar}, \mathit{cube}, \mathit{rhomb\text{-}nut}, \mathit{rim}, \mathit{tire}, \mathit{socket},\\ &\mathit{flat\text{-}washer}, \mathit{thick\text{-}washer}, \mathit{round\text{-}headed\text{-}bolt}, \mathit{hexagonal\text{-}bolt}\}\end{aligned}$$

Nine baufix®color classes used for pixel classification during the segmentation step define the labels of the variables $\boldsymbol{VColor}_i$:

$$\boldsymbol{VColor}_i \in \{\mathit{wooden}, \mathit{red}, \mathit{yellow}, \mathit{blue}, \mathit{green}, \mathit{orange}, \mathit{white}, \mathit{purple}, \mathit{ivory}\}$$

An elementary object type is uniquely defined by its type and color. A cube can have four different colors, and bolts can have five different colors[8]. All other object types have an

[7] The states of the variables are given in German language because the English translations may have a slightly different meaning. The English translations are given in Fig. 4.19 (***SType***) and Fig. 4.2 (***SColor, SShape, SSize***).

[8] The length of the thread of a bolt is coded by the color. The red bolt is the shortest, the green bolt is the longest.

individual color. All together there are 23 different elementary baufix®objects. These ***unique types*** are the possible labels of the other variables:

$$\boldsymbol{VObject_i}, \boldsymbol{VScene, IntendedObject} \in \{\textit{3-holed-bar}, \textit{5-holed-bar}, \textit{7-holed-bar}, \\ \textit{red/yellow/blue/green-cube}, \textit{rhomb-nut}, \textit{rim}, \textit{tire}, \textit{socket}, \\ \textit{flat-washer}, \textit{thick-washer}, \textit{red/yellow/orange/blue/green-round-headed-bolt}, \\ \textit{red/yellow/orange/blue/green-hexagonal-bolt}\}$$

VObject*$_i$*, $i = 1 \ldots n$, describe the unique types of the n scene objects. Given the value of this variable for an object i, the detected type and color in the image become independent random variables. The conditional probability table (CPT) $M_{VType|O}$ models the object recognizer as a statistical process. If the unique type is known to be u, $M_{VType=t|O=u}$ is the probability that the elementary object recognizer classifies t and $M_{VColor=c|O=u}$ is the probability that the pixel classifier detects color c. The first CPT has been estimated using a training set of 11 images with 156 objects:

$$M_{VType=t|O=u} = \frac{\text{\#type } t \text{ was detected when an object with unique type } u \text{ was shown}}{\text{\#objects with unique type } u \text{ that were shown}} \tag{4.16}$$

The second CPT was estimated using the training set of the pixel classifier which consisted of 27 images from 9 different scenes, with three pictures per scene:

$$M_{VColor=c|O=u} = \frac{\text{\#pixel with detected color } c \text{ when an object with color } c_u \text{ was shown}}{\text{\#pixel of the shown object with color } c_u} \tag{4.17}$$

where c_u is the color of the unique object type u.

The variable ***VScene*** summarizes which object types exist in the scene. The CPT $P(\boldsymbol{VScene}|\boldsymbol{VObject}_1, \ldots, \boldsymbol{VObject}_n)$ is treated as a noisy-or (see Sec. 3.2.3). It is constructed from the CPTs $P(\boldsymbol{VObject}_i|\boldsymbol{VScene})$. These are set to matrix I_p which is defined as follows:

$$I_p[i,j] = \begin{cases} \alpha & \text{,if } i = j, \\ \varepsilon & \text{,otherwise} \end{cases} \quad \text{,where } \alpha \text{ is near one and } \varepsilon \text{ is near zero.}$$

The variable ***IntendedObject*** denotes the unique object type that was denoted by the verbal description. It is assumed that this type will be one of the types that are present in the scene (***VScene***). The CPTs between the ***IntendedObject*** and the evidential nodes on the speech side have been estimated using the data collected in experiment 4 (WWW): $M_{SSize|O}$, $M_{SShape|O}$, and experiment 5 (Select-Obj): $M_{SType|O}$, $M_{SColor|O}$.

$$M_{F=f|O=u} = \frac{\text{\#feature } f \text{ was named for unique object type } u}{\text{\#unique object type } u \text{ was shown}}$$

where $F \in \{\boldsymbol{SType}, \boldsymbol{SColor}, \boldsymbol{SShape}, \boldsymbol{SSize}\}$.

The evaluation procedure of the Bayesian network calculates a likelihood η_i for each object i in the scene. It rates the hypothesis that object i is denoted by the verbal description, i.e. by the evidences collected in ***SType***, ***SColor***, ***SShape***, ***SSize***:

$$\eta_i = \max_j((\tau_j)_i) - \max_j((\text{offset}_j)_i) \tag{4.18}$$

$$\begin{aligned} \text{where} \quad & (\tau_j)_i = P(\boldsymbol{VObject}_i = u_j | \boldsymbol{VType}_i, \boldsymbol{VColor}_i, \boldsymbol{SType}, \boldsymbol{SColor}, \boldsymbol{SShape}, \boldsymbol{SSize}), \\ & (\text{offset}_j)_i = P(\boldsymbol{VObject}_i = u_j | \boldsymbol{VType}_i, \boldsymbol{VColor}_i), \\ & u_j \text{ is the } j\text{-th unique object type.} \end{aligned}$$

The selection criterion is defined by considering the mean μ and the standard deviation σ of all likelihood values η_i:

$$\text{object } i \text{ is selected if } \begin{cases} \sigma < \text{threshold and } \eta_i > 0 \\ \sigma \geq \text{threshold and } \eta_i > \mu + \sigma. \end{cases} \tag{4.19}$$

The second step checks the applicability of spatial relations if these have been specified by the speaker. Otherwise, the selected hypotheses from the first step are the result of the object identification procedure. For all IO/RO candidate pairs a qualitative spatial representation $\vec{\delta}(IO, RO, RF)$ is calculated. The components denote the applicability of the six projective relations. The degree of applicability is computed using the 3-d spatial model of Fuhr et al. (see Sec. 4.4.1):

$$\begin{aligned} & \vec{\delta}(IO, RO, RF)[r] = \delta(r, IO, RO, RF) \\ & \text{where } r \in \{\text{left}, \text{right}, \text{in-front-of}, \text{behind}, \text{above}, \text{below}\} \end{aligned} \tag{4.20}$$

The spatial relation that was uttered by the speaker is represented as:

$$\begin{aligned} & \vec{\rho}[r] = \begin{cases} 1.0, & \text{if } r \text{ was uttered} \\ 0.0, & \text{otherwise} \end{cases} \\ & \text{where } r \in \{\text{left}, \text{right}, \text{in-front-of}, \text{behind}, \text{above}, \text{below}\}. \end{aligned} \tag{4.21}$$

The final selection criterion s for the identification of the intended object IO and the reference object RO combines the likelihoods η_{IO}, η_{RO} of the first step and a likelihood of the spatial match:

$$s = \frac{\eta_{IO} \cdot \eta_{RO}}{||\vec{\delta}(IO, RO, RF) - \vec{\rho}||_2}, \qquad \text{where } ||.||_2 \text{ is the Euclidean distance.} \tag{4.22}$$

The candidate pair with the greatest s is selected.

4.5.2 Starting points for improvements

The work of Socher et al. was a first step towards an integrated processing of speech and images. The work presented in this thesis continues this work. It aims at improving the following aspects of the approach described in the previous subsection:

- The identification approach of Socher et al. combines different measurements and selection strategies in a hybrid manner:

  1. The object recognition and verbal description processes are modeled by a Bayesian network.
  2. The first selection step employs a statistical analysis of Bayesian network properties assuming a Gaussian distribution of likelihood values.
  3. Spatial relations are represented as fuzzified membership functions.
  4. Spatial relations are compared by an Euclidean distance measure.
  5. Spatial and type information are combined by simple multiplication rules.

  Therefore, it is very difficult to control the different parameters and threshold values. Any extension of the model will result in more specialized measurements and selection strategies.

- The two-step strategy introduces a hard decision that is only based on partial information. The redundancy that is often introduced with a spatial relation cannot be fully exploited.

- The 3-d spatial model depends on a 3-d shape reconstruction. Consequently, its accuracy is highly affected by object recognition errors that result in erroneous shape reconstructions. Objects that have been detected but classified as an unknown objects are only represented on a blob level and are therefore difficult to process on a 3-d level.

- The Bayesian modeling of verbal descriptions is a very flexible and powerful approach. However, the simple tree structure ignores some dependencies that were found in the data (see experiment 4). Furthermore, it does not take account of out-of-scope descriptions[9] and ignores the non-exclusivity of values[10]. Additionally, the speech variables of the network are only related to the whole scene – not to a single object.

- The previous work did not take into account assembled objects, their structural descriptions, and methonymian naming (e.g. *’plane’*).

The main improvements that are proposed in this thesis are the realization of a tighter interaction of vision and speech processing on the 2-d blob level and a theoretically well founded integration framework that is based on a unique Bayesian network.

[9] For example, a ’green bar’ does not exist in the domain. Nevertheless, the proposed Bayesian model is forced to explain the invalid description by a valid object class.

[10] The features *’angular’* and *’flat’* are modeled as values of the same variable ***SShape*** even though the description *”the flat angular object”* is a valid description of a bar.

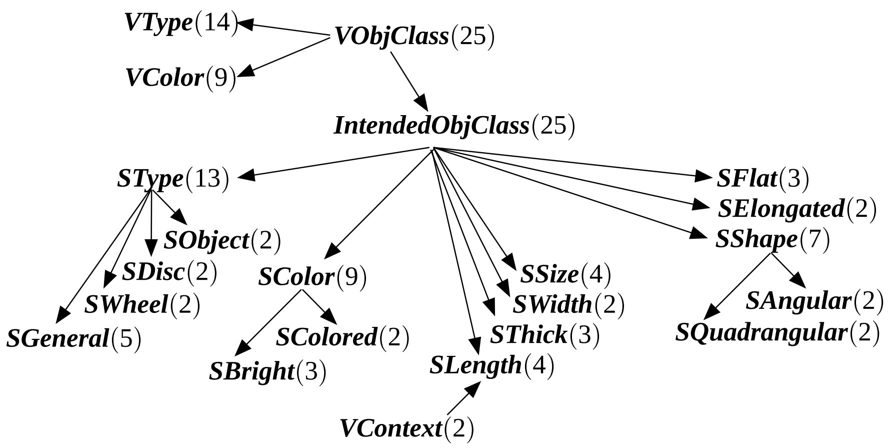


Figure 4.18: A Bayesian network modeling object classes

4.5.3 An extended Bayesian model for object classes

The modeling of corresponding variables in Bayesian networks that was proposed in the previous chapter permits us to directly connect the visual and verbal evidences of a single object (Fig. 4.18) that will subsequently be called *intended object* [WS99]. The arc from the ***IntendedObjClass*** to the ***VObjClass*** certainly depends on the value of the selection variable. This dependence will be ignored for the moment. If an utterance refers to more than one object the Bayesian network in Fig. 4.18 is built up for all specified objects. The evidential variables in this network are not only leaves because some nouns and adjectives are more precise than others. They are defined as follows:

- visual evidences:

$$\begin{aligned}
\boldsymbol{VType}_i &\in \{\textit{3-holed-bar}, \textit{5-holed-bar}, \textit{7-holed-bar}, \textit{cube}, \textit{rhomb-nut}, \textit{rim},\\
&\quad \textit{tire}, \textit{socket}, \textit{flat-washer}, \textit{thick-washer}, \textit{round-headed-bolt},\\
&\quad \textit{hexagonal-bolt}, \textit{assembly}, \textit{unknown}\}\\
\boldsymbol{VColor}_i &\in \{\textit{wooden}, \textit{red}, \textit{yellow}, \textit{blue}, \textit{green}, \textit{orange}, \textit{white}, \textit{purple}, \textit{ivory}\}\\
\boldsymbol{VContext} &\in \{\textit{3-5-7-holed-bars}, \textit{other}\}
\end{aligned}$$

The value *’assembly’* denotes a complex object in contrast to the other elementary objects. The term *’unknown’* stands for an unclassified detected object region. Two visual contexts are distinguished: if all three types of bars have been recognized in a scene the context is set to the value *’3-5-7-holed-bars’*, otherwise it is set to the value *’other’* contexts.

- type evidences extracted from speech[11]:

$$\begin{aligned}\textbf{\textit{SType}} &\in \{\text{Dreilochleiste}, \text{Fünflochleiste}, \text{Siebenlochleiste}, \text{Würfel}, \text{Raute},\\ &\quad\ \text{Felge}, \text{Reifen}, \text{Buchse}, \text{Rundkopfschraube}, \text{Sechskantschraube},\\ &\quad\ \text{Unterlegscheibe}, \text{Aggregat}, \textit{other}\}.\\ \textbf{\textit{SGeneral}} &\in \{\text{Leiste}, \text{Schraube}, \text{Mutter}, \text{Ring}, \textit{other}\}\\ \textbf{\textit{SDisc}} &\in \{\text{Scheibe}, \textit{other}\}\\ \textbf{\textit{SWheel}} &\in \{\text{Rad}, \textit{other}\}\\ \textbf{\textit{SObject}} &\in \{\text{Objekt}, \textit{other}\}\end{aligned}$$

  The states of the variable ***SType*** include all baufix®types that can be distinguished by German nouns without considering color, size, or shape. The other variables denote type abstractions. The term ’Objekt’ (*object*) refers to elementary or complex object types. It is more frequently used for complex types because it is difficult to name them.

- color evidences extracted from speech[12]:

$$\begin{aligned}\textbf{\textit{SColor}} &\in \{\text{weiß}, \text{rot}, \text{gelb}, \text{orange}, \text{blau}, \text{grün}, \text{violett}, \text{holz}, \text{elfenbein}, \text{bunt}\}\\ \textbf{\textit{SColored}} &\in \{\text{farbig}, \textit{other}\}\\ \textbf{\textit{SBright}} &\in \{\text{hell}, \text{dunkel}, \textit{other}\}\end{aligned}$$

  The color variables are organized hierarchically. All different colors in the baufix®domain are possible states of ***SColor***. Assemblies need not have a unique color. This is represented by the state ’bunt’ (*’many-colored’*). ***SColored*** and ***SBright*** are abstractions of ***SColor***.

- size evidences extracted from speech[12]:

$$\begin{aligned}\textbf{\textit{SSize}} &\in \{\text{klein}, \text{mittelgroß}, \text{groß}, \textit{other}\}\\ \textbf{\textit{SLength}} &\in \{\text{kurz}, \text{mittellang}, \text{lang}, \textit{other}\}\\ \textbf{\textit{SThick}} &\in \{\text{dick}, \text{dünn}, \textit{other}\}\\ \textbf{\textit{SNarrow}} &\in \{\text{schmal}, \textit{other}\}\\ \textbf{\textit{SHeight}} &\in \{\text{hoch}, \text{flach}, \textit{other}\}\end{aligned}$$

[11] The states of the variables are given in German because the English translations may have a slightly different meaning. The English translation are given in Fig. 4.19.

[12] The states of the variable are given in German language because their English translations may have a slightly different meaning. Their English translations can be found in Fig. 4.2.

- shape evidences extracted from speech[13]:

$$\begin{aligned}\boldsymbol{SShape} &\in \{\text{rund}, \text{sechseckig}, \text{rautenförmig}, \text{rechteckig}, \textit{other}\text{-eckig},\\ &\qquad \textit{other}\text{-viereckig}, \textit{other}\}\\ \boldsymbol{SAngular} &\in \{\text{eckig}, \textit{other}\}\\ \boldsymbol{SQuadrangular} &\in \{\text{viereckig}, \textit{other}\}\\ \boldsymbol{SElongated} &\in \{\text{länglich}, \textit{other}\}\end{aligned}$$

The states of the variable ***SShape*** distinguish round and angular shapes. From experiment 4 (WWW) it was concluded that 'eckig' (*angular*) and 'viereckig' (*quadrangular*) were not used as super-concepts of the other angular terms. Nevertheless, a detailed examination of the data yields that they are partially used as super-concepts. Therefore, the meaning has been split into the super 'eckig'/'viereckig' terms and the '*other*-eckig'/'*other*-viereckig' terms.

- intermediate and query variables:

$$\begin{aligned}\boldsymbol{VObject_i}, \boldsymbol{IntendedObjClass} \in \{&\textit{3-holed-bar}, \textit{5-holed-bar}, \textit{7-holed-bar},\\ &\textit{red/yellow/blue/green-cube}^1, \textit{rhomb-nut}, \textit{rim}, \textit{tire}, \textit{socket}, \textit{flat-washer},\\ &\textit{thick-washer}, \textit{red/yellow/orange/blue/green-round-headed-bolt}^1,\\ &\textit{red/yellow/orange/blue/green-hexagonal-bolt}^1, \textit{assembly}, \textit{other}\}\end{aligned}$$

The states of both variables denote the unique object classes of the baufix®domain. The value *'assembly'* stands for complex objects. The belief of the state *'other'* shall detect inconsistent evidence. Either the speaker did not refer to an object in the scene, or the visual object was misclassified, or erroneous speech interpretations occurred that lead to invalid verbal object descriptions like *"the green bar"*.

The conditional probability tables (CPTs) of the *IntendedObjClass*-model are partially set by hand and partially estimated from data. On the vision side, the CPT $M_{VColor|O}$ has been taken from the previous modeling (Eq. 4.18). The CPT $M_{VType|O}$ has been estimated from the recognition results of the image set used in experiment 5 (Eq. 4.16). The ***SType*** and ***SColor*** subnets have been hand-modeled considering some qualitative results of experiment 5:

1. The ***SType*** subnet realizes the abstraction hierarchy shown in Fig. 4.19. The CPTs are defined as follows:

$$M_{X=t|SType=u} = \begin{cases} \alpha_t, & \text{if noun } t \text{ is an abstraction of noun } u \\ 0.0, & \text{otherwise} \end{cases}, \quad \sum_t \alpha_t = 1.0$$

where α_t is correlated with the strength of the relation,

$$X \in \{\textit{SGeneral, SDisc, SObject, SWheel}\}.$$

[13] The states of the variable are given in German language because their English translations may have a slightly different meaning. Their English translations can be found in Fig. 4.2.

[1] The notation *red/yellow/blue/green-cube* is a short term for *red-cube, yellow-cube, ...*

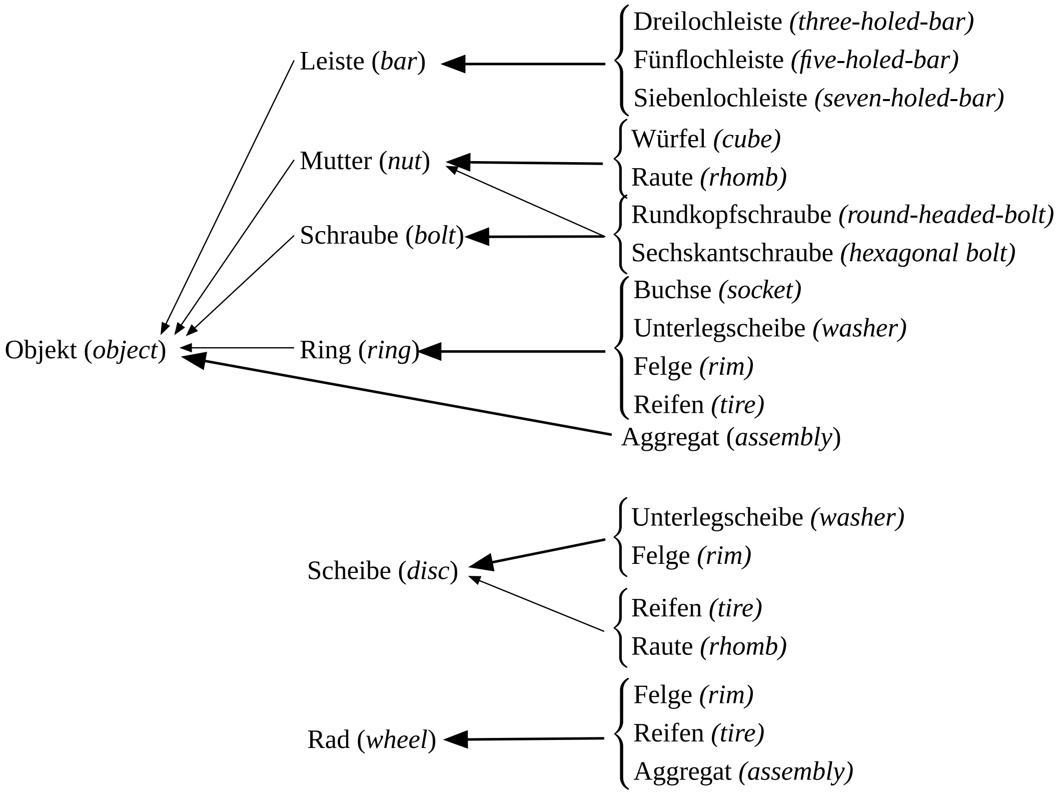


Figure 4.19: Hierarchy of object type names: the arrows indicate abstractions, the width of the arrows indicate the strength of the relationship.

2. The CPT $M_{SType|O}$ maps the object types to the unique object types, e.g.:

$$M_{SType=\text{Würfel}|O=red\text{-}cube} = 1.0$$
$$M_{SType=\text{Unterlegscheibe}|O=flat\text{-}washer} = 1.0$$
$$M_{SType=\text{Seckskantschraube}|O=blue\text{-}hexagonal\text{-}bolt} = 1.0$$

3. The CPT $M_{SColor|O}$ models the relationship of the verbally mentioned colors and the unique object types. Most elementary baufix®objects have a definite color like red, green, blue, yellow, or white. Objects with mixed colors like purple, orange, wooden, or ivory are often described by one of the definite colors: red for orange, blue for purple, white for wooden or ivory. These cases have been considered in the CPT definition.

4. The CPTs $M_{SColored|SColor}, M_{SBright|SColor}$ define abstractions of the concrete colors. They have been set based on introspection.

The CPTs of the size and shape adjectives are very difficult to set by hand. The qualitative results of experiment 4 have shown that the usage is determined by multiple causes and context dependencies. Therefore, these CPTs have been estimated directly from the data of experiment 4 (WWW):

$$M_{F=f|O=t} = \frac{\#\text{feature } f \text{ was named for unique object type } t}{\#\text{unique type } t \text{ was shown}} \tag{4.23}$$

where $F \in \{\textit{\textbf{SSize, SWidth, SThick, SFlat, SElongated, SShape}}\}$.

For the ’*other*-eckig’ (*angular*) and ’*other*-viereckig’ (*quadrangular*) states of ***SShape*** only those denotations are counted that did not co-occur with one of the other angular terms. For the *assembly* and *other* types a constant probability distribution is assumed.

The ***SAngular*** and ***SQuadrangular*** variables abstract from the states of ***SShape***. The CPTs are estimated as follows:

$$M_{F=f|SShape=s} = \frac{\#\text{feature } f \text{ and feature } s \text{ were selected for the same object}}{\#\text{feature } s \text{ was selected}} \tag{4.24}$$

where $F \in \{\textit{\textbf{SAngular, SQuadrangular}}\}$.

The usage of the adjectives ’kurz’ (*short*), ’mittellang’ (*medium-long*), and ’lang’ (*long*) greatly depends on the context in case of bars. If all three types of bars are present, the *5-holed-bar* is more frequently called ’mittellang’ than ’lang’. In other contexts, the opposite is true. In parallel, the *7-holed-bar* is more frequently called ’kurz’ in the first case. Therefore, the CPT is estimated in two different contexts:

$$M_{SLength=f|O=t,VContext=c} = \frac{\#\text{feature } f \text{ was named for unique object type } t \text{ in context } c}{\#\text{unique type } t \text{ was shown in context } c}$$

where $c \in \{\textit{3-5-7-holed-bars}, other\}$.

For each CPT that is used in the speech subnet a small offset ε is added to each conditional probability value. After that, the CPTs are normalized again. By this means, the Bayesian network takes account of speech recognition errors that may lead to inconsistent evidence.

The novel Bayesian network proposed in this subsection has a basic structure which is similar to that of Socher et al.: The ***IntendedObjClass*** d-separates the type, color, size, and shape evidences that were extracted from speech and the visual evidences of the intended object. This structure reflects the qualitative result of the WWW experiment (experim. 4) about the object-class-specific meaning of size and shape adjectives.

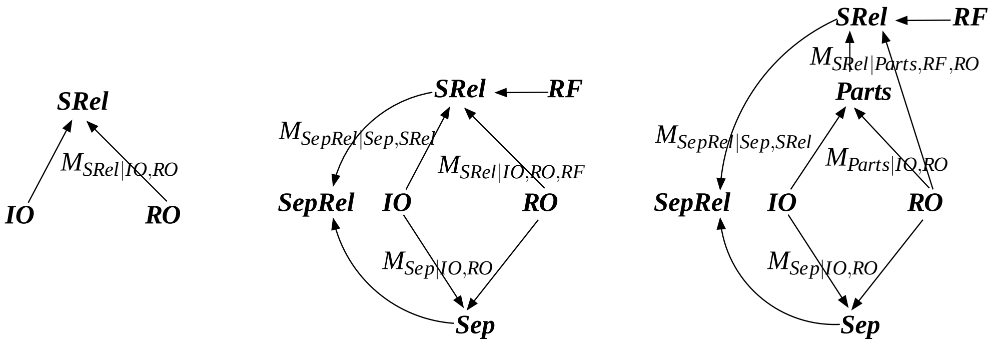
4.5.4 A Bayesian model for spatial relations

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.*”

In subsections 4.4.2,4.4.3 the interpretation of the projective binary relations *left, right, in-front-of, behind, above, below* is extensively discussed. This section will show how the spatial model can be integrated into the probabilistic framework that was proposed in the previous chapter.

The Bayesian modeling of a projective relation can be examined on different abstraction layers:



(a) Basic relation

(b) Explicit separation and reference frame

(c) Explicit partitioning

Figure 4.20: A Bayesian spatial model.

1. A black box relation between the intended object $IO = io$ and the reference object $RO = ro$. The degree of applicability is calculated with regard to a default reference frame, e.g. a speaker-centered reference frame.

2. Two coupled relations between the object areas $\mathcal{A}^{IO}, \mathcal{A}^{RO}$. The first measures the degree of applicability of the 2-d direction $\vec{d}^{RF}_{2D}$ to the object areas with regard to a selected reference frame. The second measures the degree of separation between the object areas.

3. Relations that are defined on a space partitioning $\mathcal{A}^{ro}_{k=1...n}$. The degree of containment and the degree of accordance are modeled seperately.

Bayesian networks provide the language for modeling each of these abstraction layers. The trade-off is an increased model complexity and computation time versus an increased power of the query language.

The simplest Bayesian model for projective relations is presented in Fig. 4.20(a). Let n be the number of objects that have been detected in the scene. The random variables $IO, RO \in \{1, \ldots, n\}$ denote the possible intended and reference objects. $SRel \in \{r_1, \ldots, r_5\}$ represents the possible states of the spatial relation. A state of a projective spatial relation is a vector on the 2-d image plane that defines the associated direction of the specified relation. The continuous space of direction vectors is discretized into four different directions: r_1 is the specified direction $\vec{d}^{RF}_{2D}$, $r_2 = -r_1$ is the negative direction, and $r_3 = r_1^{\perp}, r_4 = -r_1^{\perp}$ are the two possible orthogonal directions. A fifth state r_5 represents the state that none of first four states of the relation will hold due to the separation criterion or because the object area of the reference objects completely includes that of the intended object. In the simplest Bayesian model the spatial model is treated as a black box providing the CPT $M_{SRel|IO,RO}$:

$$M_{SRel=r_i|IO,RO} = \beta\, SModel(r_i, IO, RO), \tag{4.25}$$

where $SModel$ is the black box function of the spatial model,
β is a normalizing constant.

The black box function contains different weighting criteria like applicability and separation. Different reference frames are not explicitly modeled. The second more detailed version of a Bayesian network shown in Fig. 4.20(b) considers these two aspects. The variable $RF \in \{rf_1, \ldots, rf_m\}$ denotes different possible reference frames, e.g. speaker-centered vs. hearer-centered (cf. Sec. 2.5.2). The variables $Sep, SepRel \in \{true, false\}$ model the assumption that two objects which are used in a spatial relation should not be separated by other objects. The CPTs are defined as follows:

$$M_{SRel=r_i|IO=io,RO=ro,RF=rf_j} = \beta \begin{cases} \delta(r_i, io, ro, rf_j), & \text{if } i \neq 5 \\ \varepsilon, & \text{if } i = 5 \end{cases} \tag{4.26}$$

where $\delta(,)$ calculates the degree of applicability of the relation state r_i,
ε is a small constant and β is a normalizing constant.

$$M_{Sep=true|IO=io,RO=ro} = \begin{cases} 1-\varepsilon, & \text{if } Sep(\mathcal{A}_{io}, \mathcal{A}_{ro}, \{A_k | k = 1 \ldots n, k \neq io, k \neq ro\}) > \Theta_{Sep} \\ \varepsilon, & \text{otherwise} \end{cases}$$

where $Sep(,,)$ calculates the degree of separation, (4.27)
ε is a small constant.

$$M_{SepRel=true|Sep,SRel} = \begin{cases} 1-\varepsilon, & \text{if } Sep = false \wedge SRel \neq r_5 \\ \varepsilon, & \text{if } Sep = true \wedge SRel \neq r_5 \\ 0.5, & \text{otherwise} \end{cases} \tag{4.28}$$

where ε is a small constant.

Note that an elimination[14] of the variables RF, $SepRel$, and Sep results in the previous network structure.

The Bayesian network presented in Fig. 4.20(c) realizes the next level of detail. The states of the variable $Part \in \{\mathcal{A}_k^{ro} | k = 1 \ldots m_{ro}, ro = 1 \ldots n\}$ are the areas of the space partitioning scheme. The CPTs encode the degree of containment and the degree of

[14] Elimination means the summation over all states of a variable. This has been defined as one operation of the bucket elimination aligorithm in Sec. 3.3.3.

accordance:

$$M_{Part=\mathcal{A}_k^{ro'}|IO=io,RO=ro} = \begin{cases} \gamma(\mathcal{A}_k^{ro'}, \mathcal{A}^{io}), & \text{if } ro' = ro \\ 0.0, & \text{otherwise} \end{cases} \tag{4.29}$$

where $\gamma(,)$ calculates the degree of containment of object *io* in partition $\mathcal{A}_k^{ro'}$.

$$M_{SRel=r_i|Part=\mathcal{A}_k^{ro'},RO=ro,RF=rf_j} = \beta \begin{cases} \alpha(\vec{d}_k^{ro'}, \vec{d}_{r_i}^{rf_j}), & \text{if } ro' = ro \wedge i = 1\dots4 \\ \varepsilon, & \text{if } ro' = ro \wedge i = 5 \\ 1, & \text{otherwise} \end{cases} \tag{4.30}$$

where $\vec{d}_k^{ro'}$ is the direction vector that is associated with the k-th partition of the reference object ro',

$\vec{d}_{r_i}^{rf_j}$ is the direction of the i-th state of *SRel* with regard to the reference frame rf_i,

$\alpha(,)$ calculates the degree of accordance of the direction r_i and the partition $\mathcal{A}_k^{ro'}$.

ε is a small constant and β is a normalizing constant.

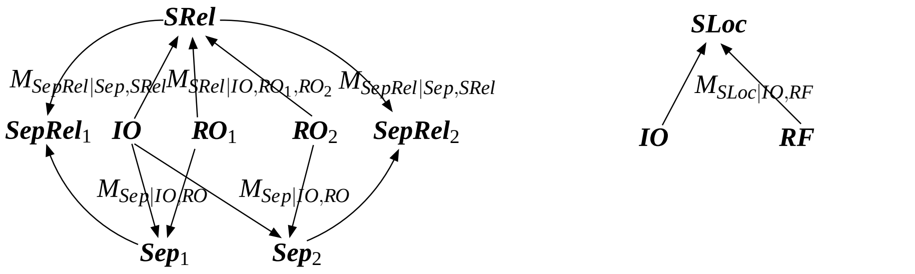
Again, the elimination of the variable *Part* results in the previously described Bayesian network.

The more detailed the Bayesian network is, the more flexible is the usage of the spatial model:

1. The simplest model (Fig. 4.20(a)) can only answer the queries which objects fulfill the constraints of the specified relation or which state of a relation is most probable if the intended and reference objects are given. The reference frame is fixed by default.

2. The more complex network shown in Fig. 4.20(b) provides the possibility to explicitly select one of different possible reference frames, to specify a priori probabilities for the selection of reference frames, or even to query the reference frame selected by a speaker. The explicit modeling of the separation property of spatial relations can also be exploited. The detection that two objects are separated basically depends on the decomposition of an image into foreground (objects on the table) and background (table plane). If this decomposition is not straightforward the most probable state of this variable can be used in order to decide whether other objects are placed between two specified objects or not.

3. In the first two networks, a spatial relation must be anchored by an intended object and a reference object. The area that may be denoted by a specified spatial relation cannot be queried, e.g. *”Place the object in front of the blue cube”*. In the third Bayesian network (Fig. 4.20(c)) such a query can be realized by extending the set of states of the variable $IO \in \{1,\dots,n,\perp\}$. In the corresponding components of the



(a) Trinary relation between

(b) Locative prepositions

Figure 4.21: Bayesian networks for the spatial relation *between* and locative prepositions such as *on-the-right*.

CPT, the areas of the different space partitioning are equally probable:

$$M_{Part=\mathcal{A}_k^{ro'}|IO=\perp,RO=ro} = \beta \begin{cases} 1, & \text{if } ro' = ro \\ 0, & \text{otherwise} \end{cases} \tag{4.31}$$

where β is a normalizing constant.

Querying the variables $RO, Part$ yields the most probable area where an object should be located or placed.

In the following result chapter, the second Bayesian network is used for most experiments. As long as an intended object that is detected by the vision components should be identified, an explicit modeling of the space partitioning is not needed.

Other descriptions of the position of objects can be modeled by similar Bayesian networks. Fig. 4.21 presents the network structures of the relation *between* and locative prepositions such as *on-the-left*.

In the *between* network the states of the variable ***SRel*** are defined relatively to the direction between the intended object $IO = io$ and the second reference object $RO_2 = ro_2$. The definition of the CPT $M_{SRel|IO,RO_1,RO_2}$ is based on the degree of accordance of state r_i and the direction between io and ro_1.

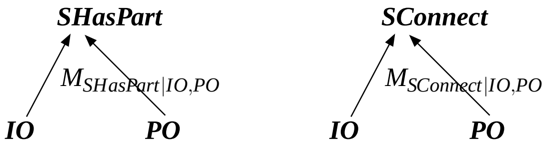
The variable ***SLoc*** in the network for locative prepositions has two possible values indicating whether the preposition is valid or not. The CPT values $M_{SLoc=true|IO=io,RF=rf_i}$ are defined by the corresponding degree of applicability.

4.5.5 Modeling structural relationships

Complex baufix®objects consist of elementary objects that have been screwed or plugged together. They can be verbally described by enumerating the elementary subparts:

"[...] the object consisting of the bar and the cube."

Elementary objects that are part of an assembly can be described by specifying parts that are connected to them:



(a) has-part relation

(b) connected relation

Figure 4.22: Bayesian networks for structural relationships.

”[. . .] the cube with the bolt.”

Indeed, the semantics of the relation *with* is ambiguous. The bolt may either be connected with the cube or be part of it: *”[. . .] the airplane with the red cockpit.”* On the vision side the recognition module for complex objects extracts assembly structures that define which elementary objects are connected by mounting relations. These structures are used in order to establish the CPTs of the *has-part* and *connected* relations:

$$M_{SHasPart=true|IO=io,PO=po} = \begin{cases} 1-\varepsilon, & \text{if object } po \text{ is part of a complex object } io \\ \varepsilon, & \text{if the object area of } po \text{ touches that of the complex area } io \\ 0, & \text{otherwise} \end{cases} \tag{4.32}$$

$$M_{SConnect=true|IO=io,PO=po} = \begin{cases} 1-\varepsilon, & \text{if objects } io, po \text{ are part of the same complex object} \\ 1-\varepsilon, & \text{if object } po \text{ is part of a complex object } io \\ \varepsilon, & \text{if objects areas of } io, po \text{ touch each other} \\ 0, & \text{otherwise} \end{cases} \tag{4.33}$$

where ε is a small constant.

The CPT definitions consider the uncertainty of the recognition module for complex objects by introducing the small constant ε. Erroneous results of this module are mainly consequences of propagated errors from the elementary object recognition module.

4.5.6 Integrating the what and where

In the previous subsections the components of the integration model were presented. How these subnets interact during an inference process shall be discussed by means of the following example:

Example 6 *Four objects are placed on a table (see Fig. 4.23): a flat wooden washer, a thick purple washer, a red rim, and a complex object consisting of a red cube and two bars that are fixed to the cube by means of a blue bolt. The speaker instructs the system with the sentence:*

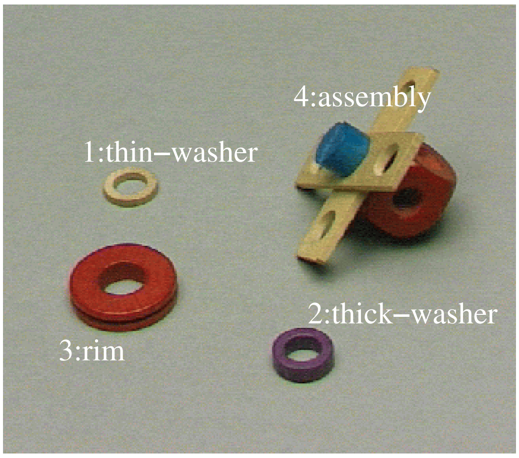


Figure 4.23: Example of a baufix® scene.

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

The system should figure out which object is intended and confirms the instruction by the answer:

System: *"Ok. Ich nehme die lila Unterlegscheibe."*
[Ok. I will take the purple washer.]

In order to simplify the discussion of the integration network, it is assumed that no object recognition errors or erroneous speech interpretations occurred (Fig. 4.24). The more general case will be examined in the next chapter. The object class and spatial subnets are integrated through the selection variables IO, RO_1. The evaluation of the Bayesian network is started by determining the maximum a posteriori hypothesis (map) of the variables IO, RO_1:

$$(io^*, ro^*) = \operatorname*{argmax}_{io,ro} P(IO = io, RO_1 = ro|\mathcal{E}) \tag{4.34}$$

where $\mathcal{E}$ is the set of speech and vision evidences.

The objects $IO = 2, RO_1 = 4$ will be identified based on the spatial arrangement of the scene objects. In order to generate the statement from example 6 a next step has to be performed that generates a more precise description of the intended object:

$$(typ^*, col^*) = \operatorname*{argmax}_{typ,col} P(SType_0 = typ, SColor_0 = col|IO = io^*, \mathcal{E}) \tag{4.35}$$

The speech evidences $SGeneral_0$ =*ring*, $SSize_0$ =*small* and the vision evidences $VType_2$ =*thick-washer*, $VColor_2$ =*violet* yield the type *washer* and the color *purple* as the maximum a posteriori hypotheses.

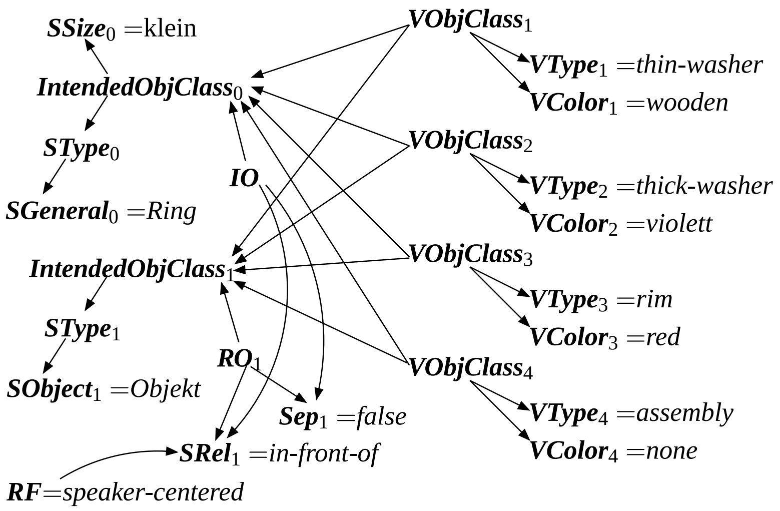


Figure 4.24: Bayesian network for integrating the speech and vision evidences of example 6. Evidential variables that can be eliminated without influencing the belief of the remaining variables are ignored.

4.6 Summary

This chapter presented the ***decoding framework*** (cf. Postulate 1) for the solution of the ***correspondence problem*** (cf. Sec. 2.5). The object recognition and speech understanding parts are interpreted as stochastic processes that are determined by various kinds of uncertainty. The probabilistic model explicitly takes account of

- erroneous classification results of the object recognizer,
- undetected mounting relations,
- the lexical ambiguity of words describing the type, color, size, and shape of an object,
- and the referential uncertainty that is introduced by spatial descriptions.

It implicitly takes account to speech recognition errors, as well for syntactic and semantic ambiguities by inference processes which will be discussed in the next chapter.

The proposed Bayesian network considers object descriptions on different ***levels of abstraction***. On the vision side, the region segmentation module provides a first representation of an object hypothesis on a blob level that may be extended and specified by the object recognition component. Both kinds of hypotheses can be processed using the integration network. On the speech side, a hierarchy of type names and adjectives has been modeled that represent different granularities of verbal descriptions. The 3-d spatial model that was proposed in the previous work of Socher et al. was transferred to the simplest abstraction layer – the 2-d blob level. By this means, the redundancy that is often introduced by spatial relations can even be exploited for objects that are only represented on a blob level.

The ***conditional probability tables*** contain the parameters of the model. They are estimated from data collected in psycholinguistic experiments (experim. 4), calculated from computational models (spatial model, complex object recognition), or defined by hand. The hand-modeled conditional probabilities and the structure of the Bayesian network take account of qualitative results that have been observed in different experiments (Experim. 4, 5).

The integration component is realized as an ***independent active interaction component*** (Postulate 2). The vision and speech understanding modules are considered as black boxes providing information on different levels of abstraction. Based on the intensional model presented in this chapter, different inference processes can be defined that realize the interaction between speech and vision components. These will be examined and discussed in more detail in the next chapter.

Chapter 5

Inference and Learning

In the last chapter the integration model was presented that is used in order to establish referential links between visual and verbal representations. The correspondence problem has been solved by finding the most probable hypothesis that explains the evidence instantiated in the Bayesian network. Further more 5 different inference tasks have been identified that perform different interactions between the separate speech and image interpretation components. The realization of these tasks in the proposed integration network is the topic of this chapter. Each task will be illustrated by performance examples that show the effectiveness of the model.

The tasks either perform a disambiguation or an enrichment of a visual or verbal representation. This is a prerequisite for learning. Before the system can learn new categorical concepts strategies must be implemented that learn new facts about a particular scene. Such strategies are realized by probabilistic inferences in the proposed Bayesian network. These inference may not be truth preserving. The most probable states of random variables can change if new evidences are considered. Therefore, it is related to ***inductive learning*** (cf. Sec. 2.5.3). The feedback of a learning step is given by the human speaker in a dialog: the way he will react on the action of the system. Consequently, the situation of the system is that of ***reinforcement learning***.

5.1 Establishing referential links

The first step in performing any inference task is the identification of the intended object (task 2, ***mp-objs***). The simplest case occurs if the intended object is described directly without considering other reference objects:

$$(io^*) = \operatorname*{argmax}_{io} P(IO = io|\mathcal{E}) \tag{5.1}$$

The probability $P(IO = io^*|\mathcal{E})$ can be interpreted as the ***plausibility*** η_{io^*} of object io^* to be denoted by the verbal object description. This plausibility can be calculated for each possible object in the scene. If a reference object was mentioned by the speaker the

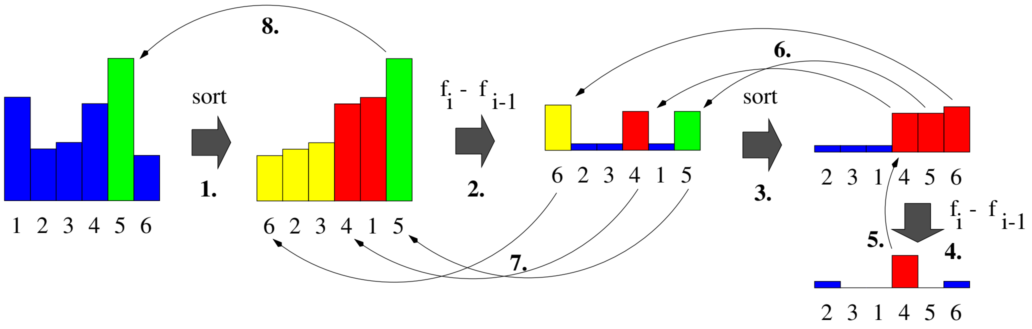


Figure 5.1: Selection of the group of possibly intended objects.

plausibility is defined by:

$$\eta_{io} = \alpha \max_{ro} P(IO = io, RO = ro|\mathcal{E}), \quad \text{where } \alpha \text{ is a normalizing constant.} \tag{5.2}$$

In equation 5.1 the maximum operation is used in order to select a particular denoted object. In many cases such a selection cannot be definite because of several reasons:

- The speaker intended to specify a group of objects.
- The speaker did not realize that the naming was not specific for the object in mind.
- Some attributes mentioned by the speaker were misrecognized or not recognized at all. Some descriptions may be misinterpreted.
- Some objects in the scene may be misrecognized due to segmentation of classification errors.

Consequently, instead of a single object a group of objects has to be selected. The remaining referential ambiguity can be resolved by the dialog component, i.e. by querying more specific information and thereby increasing the redundancy of the verbal description.

The selection of the group of possibly intended objects is based on the ***plausibility vector*** $\vec{\eta}$:

$$\vec{\eta} = (\eta_1, \ldots, \eta_n), \quad \text{where } \eta_{io} = P(IO = io|\mathcal{E}) \tag{5.3}$$

This vector shall be partitioned into one group of components that defines the query answer and one group of components that denote the objects not queried. The selection of an appropriate threshold is very difficult because no information about the distribution of plausibility values is given. Intuitively, the selection of the most plausible group of objects will be based on an examination of the differences between the components of the plausibility vector. For example, if the maximum component of the vector is significantly higher than the next greatest component, only the maximum should be selected. If three components of the vector are significantly higher, these should be selected. The following algorithm realizes such a scheme in a dynamic and flexible way [WBPS$^+$99, WBPK$^+$99]. The processing steps (1.) to (7.) are visualized in Fig. 5.1:

1. In order to calculate the differences between the components, these are sorted:

$$\tilde{\vec{\eta}} = sort(\vec{\eta})$$

2. Any zero components of the vector are ignored in the further processing steps. Thus, the minimum non-zero component $\tilde{\eta}_j$ has no valid difference value. Therefore, a parameter β is introduced to the algorithm that defines this value between 0 and $\tilde{\eta}_j$. In the following it will be called ***zero-partitioning line***:

$$\Delta^{\tilde{\eta}} = (\tilde{\eta}_j - \beta \cdot \tilde{\eta}_j, \tilde{\eta}_{j+1} - \tilde{\eta}_j, \ldots, \tilde{\eta}_n - \tilde{\eta}_{n-1}),$$

where j is the index of the first non-zero component
β is the *zero-partitioning line*.

3. The difference vector $\Delta^{\tilde{\eta}}$ may have more than one significant component value (see Fig. 5.1, step (2.)). Therefore, the same technique is applied to this vector in order to select the most significant difference values, i.e. sorting the difference vector

$$\tilde{\Delta}^{\tilde{\eta}} = sort(\Delta^{\tilde{\eta}})$$

4. and calculating the component difference:

$$\Delta^{\tilde{\Delta}^{\tilde{\eta}}} = (0, \tilde{\Delta}_2^{\tilde{\eta}} - \tilde{\Delta}_1^{\tilde{\eta}}, \ldots, \tilde{\Delta}_n^{\tilde{\eta}} - \tilde{\Delta}_{n-1}^{\tilde{\eta}})$$

5. This time, only the most significant value is relevant because the difference vector $\Delta^{\tilde{\eta}}$ shall only be partitioned into significant and non-significant differences:

$$i^* = \underset{i}{\operatorname{argmax}} \Delta_i^{\tilde{\Delta}^{\tilde{\eta}}}$$

At this point, the algorithm decides which components of the plausibility vector are selected. The next processing steps propagate this decision back through the previous transformations (see Fig. 5.1, steps (5.), (6.), (7.), (8.)).

6. Select the most significant differences:

$$\mathcal{J} = \{j_k | \Delta_{j_k}^{\tilde{\eta}} = \tilde{\Delta}_i^{\tilde{\eta}} \wedge i \geq i^*\}$$

7. Select the most significant values:

$$\mathcal{H} = \{h_l | \eta_{h_l} = \tilde{\eta}_j \wedge j \geq \max\{j_k | j_k \in \mathcal{J}\}\}$$

The ***zero-partitioning line*** automatically adapts during several dialog steps. In Fig. 5.2 the algorithm selected three components of the plausibility vector that correspond to three different objects in the scene. In the next dialog step the speaker may just repeat the instruction, e.g. *”The long one.”*. The values of the three components of the plausibility vector may remain the same, but the zero-partitioning line has been increased because any other component is set to zero by considering the dialog context. This time only component 5 is selected, and a precise answer is returned by the system.

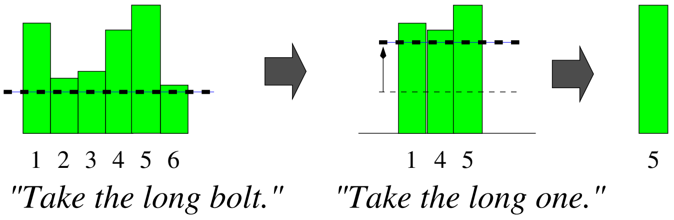


Figure 5.2: Influence of the ***zero-partitioning line***: the parameter β is defined relative to the minimum non-zero component. By this mean, it automatically adapts during successive dialog steps.

5.2 Interaction of speech and image understanding

If the verbal description has been linked with a unique visual object further inferences can be performed. In the next subsections these will be illustrated by several performance examples that have been partially taken from the evaluation set 5 (Select-Obj).

5.2.1 The most probable class of the intended object

If a speaker instructs the system to perform an action with an intended object, the class of the object plays an important rule because it determines this or successive actions that have to be performed. The proposed integration component is able to figure out the intended object *and* class despite vague descriptions of the speaker or erroneous recognition results by the speech or vision components (task 3, ***mp-class***) [WFKS00].

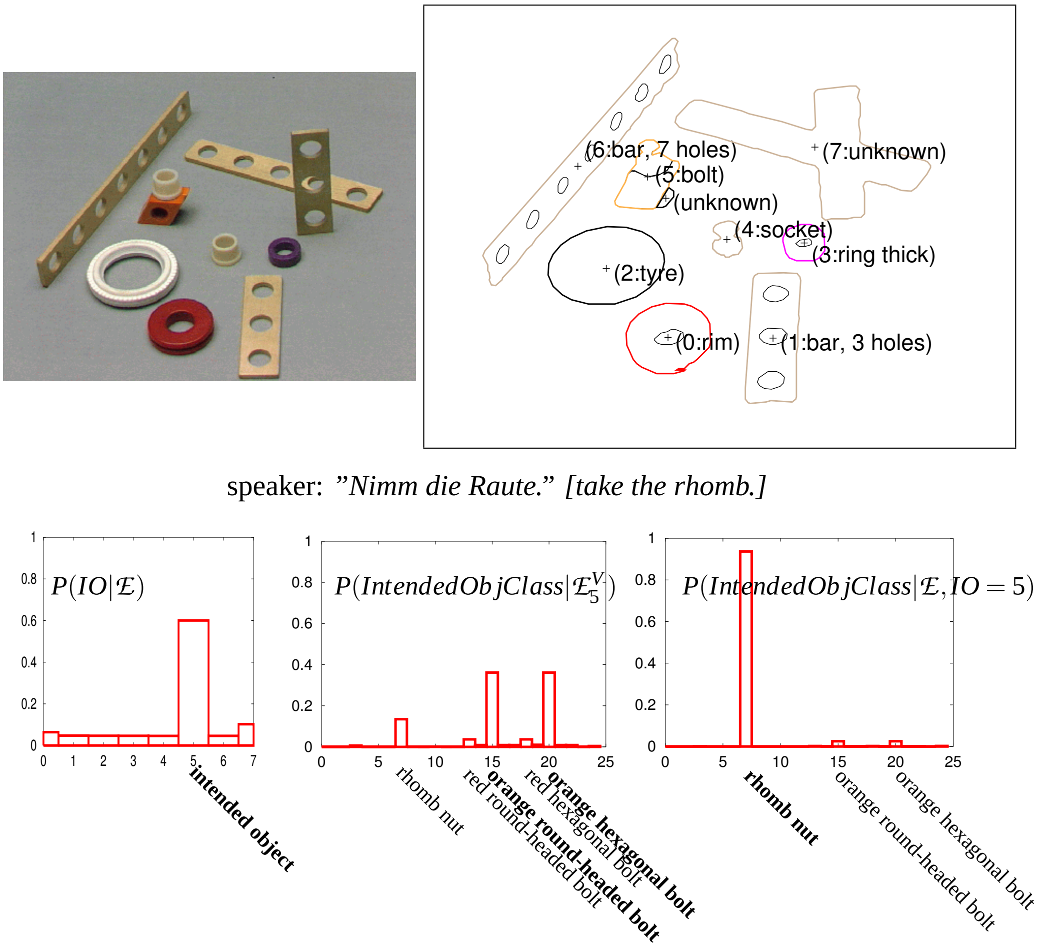
Performance example 1 is shown in Fig. 5.3. The *rhomb-nut* and the *socket* have been misclassified as the head and the thread of an *orange bolt*. The speaker sees the rhomb-nut in the scene and tells the system to take it: *"Take the rhomb."*

The intended object is precisely specified in the verbal description, but erroneously in the visual description. The most probable object class is calculated by the following equations:

$$\begin{aligned} o &= \underset{o}{\operatorname{argmax}} P(IntendedObjClass = o | IO = io^*, \mathcal{E}) \qquad (5.4) \\ \text{where } (io^*) &= \underset{io,ro}{\operatorname{argmax}} P(IO = io | \mathcal{E}). \end{aligned}$$

In Fig. 5.3 object 5 has the maximum plausibility and is the only one selected. The next two graphs show the state change of the variable ***IntendedObjClass*** if the verbal evidence is additionally considered. The correct object class *rhomb-nut* has been inferred.

Performance example 2 has been calculated on the same visual scene as the first one. This time the *5-holed-bar* that is partially occluded by the *3-holed-bar* is the intended object: *"Take the five-holed bar."*



speaker: *"Nimm die Raute." [take the rhomb.]*

Figure 5.3: Performance example 1: The speaker refers to object 5. This object has been incorrectly classified: $\mathcal{E}_5^V = \{orange, bolt\}$. The evidence extracted from speech is $\mathcal{E}^S = \{\text{Raute}\}$. $\mathcal{E} = \bigcup_{i=0\ldots7} \mathcal{E}_i^V \cup \mathcal{E}^S$. The system correctly selects object 5 and infers the correct object class.

Again, the five-holed bar has not been recognized. The segmentation algorithm generated a common wooden region for both overlapping bars in the scene. The object recognizer classified this object region as an *unknown* object. In Fig. 5.4 the plausibility vector $\vec{\eta} = P(IO|\mathcal{E})$ is presented. The seven-holed bar (obj. 6) and the unknown object (obj. 7) are selected. Now, it is up to the human communication partner to chose the correct one, e.g. by specifying a spatial relation. The seven-holed bar has been selected because *5-holed-bars* are frequently misrecognized as *7-holed-bars*. The detected *unknown, wooden* object has been selected because its correct class may be one of several possible classes as shown in the second graph of Fig. 5.4. If the verbal evidence is considered for the unknown wooden object the correct object type *5-holed-bar* is inferred (third graph in Fig. 5.4).

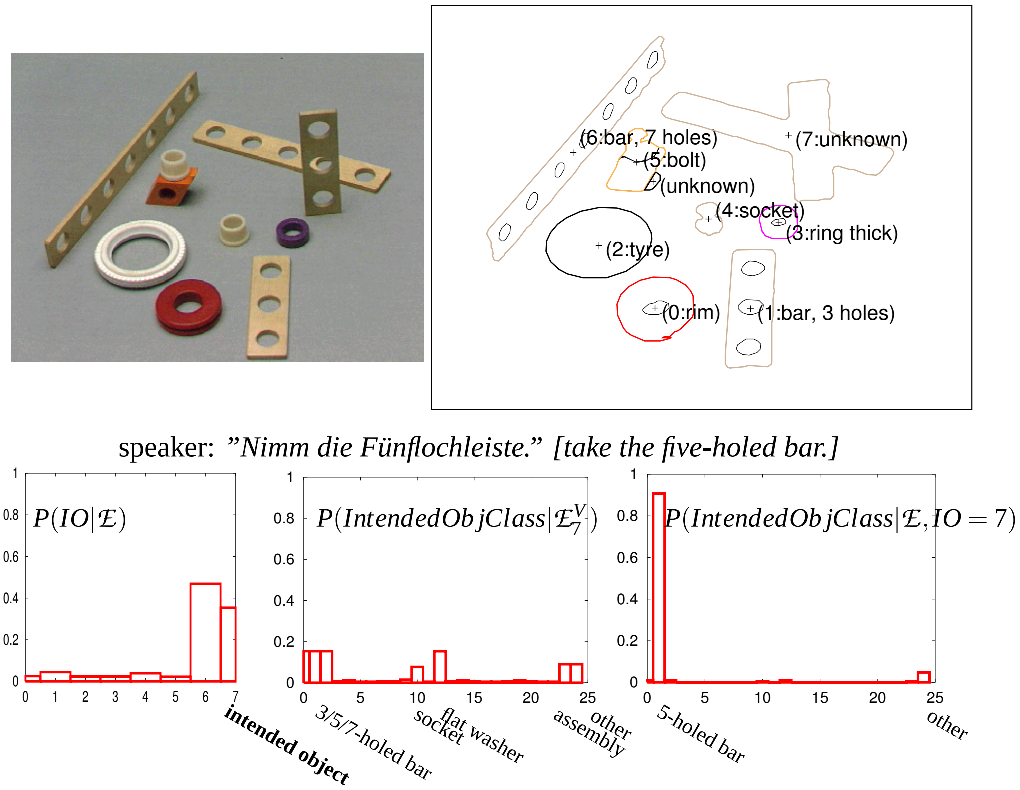
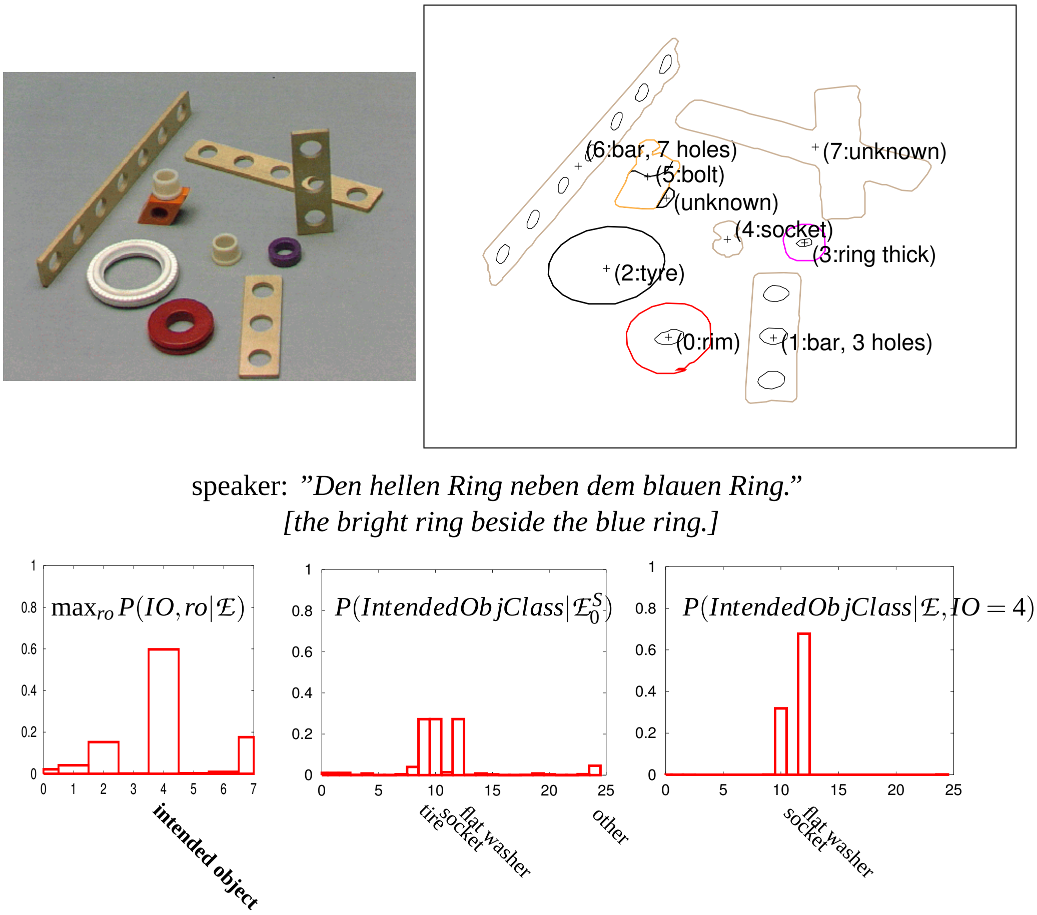


Figure 5.4: Performance example 2: The visual scene is that of Fig. 5.3. The speaker refers to object 7. It is only detected as an unknown region: $\mathcal{E}^V_7 = \{wooden, unkown\}$. The evidence extracted from speech is $\mathcal{E}^S = \{\text{Fünflochleiste}\}$. $\mathcal{E} = \bigcup_{i=0...7} \mathcal{E}^V_i \cup \mathcal{E}^S$. The system selects objects 6 and 7. For object 7 the correct class is inferred.

Performance example 3 demonstrates the processing of a more complex verbal object description: *”The bright ring beside the blue ring.”*

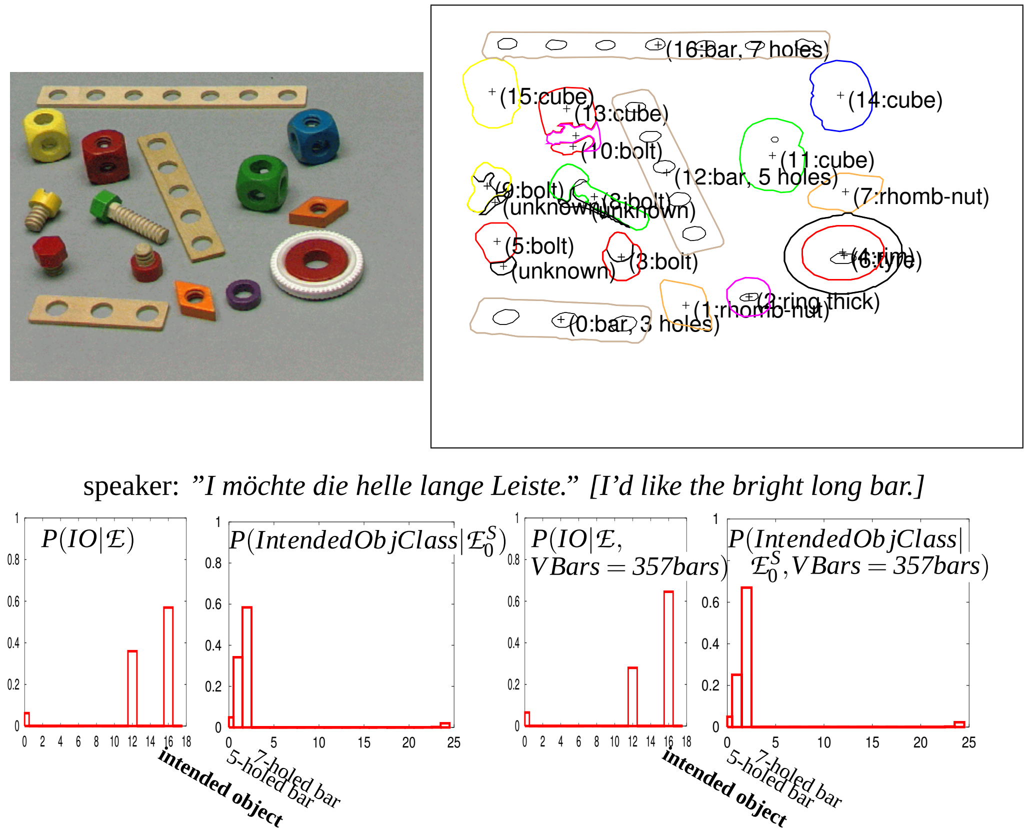
The visual context is that of the previous examples. This time the object type is only described by vague attributes. The second graph in Fig. 5.5 shows the distribution of object classes that could be intended. The spatial relation *beside* introduces a neighborhood relation between the *blue ring* and the *bright ring*. Thus, the white tire is excluded and the *socket* (obj. 4) is correctly selected (first graph in Fig. 5.5). The third graph in Fig. 5.5 shows the belief of the ***IntendedObjClass*** if the whole evidence is considered. The most probable class is that of a *wooden flat-washer*, the second one is the *ivory socket*. This is an artifact of the system. *Flat-washers* have not been modeled in the visual component of the system. Therefore, the object recognizer classifies each *flat-washer* that is present in a scene to the most similar object class, i.e. *socket*. This is reflected by the conditional probabilities in the Bayesian network. Thus, although the visual component has correctly classified the type *socket* the system infers a *flat-washer* that is additionally supported by the classified *wooden* color of the object region. This is an interesting example how the system can deal with only partially modeled object classes.



speaker: *"Den hellen Ring neben dem blauen Ring."*
[the bright ring beside the blue ring.]

Figure 5.5: Performance example 3: The visual scene is that of Fig. 5.3. The speaker refers to object 4 and reference object 3. Both object types have been correctly recognized: $\mathcal{E}_4^V = \{wooden, socket\}$ (incorrect color), $\mathcal{E}_3^V = \{purple, washer\}$. The evidence extracted from speech is $\mathcal{E}_0^S = \{\text{hell, Ring}\}$, $\mathcal{E}_1^S = \{\text{neben, blau, Ring}\}$. $\mathcal{E} = \bigcup_{i=0\ldots7} \mathcal{E}_i^V \cup \mathcal{E}_0^S \cup \mathcal{E}_1^S$. The system correctly selects the object 4 and reference object 3. The most probable object class is *flat-washer*.

Performance example 4 shows the influence of the general visual scene context. The utterance *"I'd like the bright long bar."* describes the intended object class by the attribute *long* which semantics are context dependent. In the general case the *5-holed-bar* and the *7-holed-bar* are called *long* with a high frequency. However if all three types of bars are present an ordering is introduced to the bar classes in the scene. They are mentally sorted by length. The *3-holed-bar* is the *short*, the *5-holed-bar* is the *middle-long*, and the *7-holed-bar* is the *long* one. These facts are summarized by the estimated conditional probability tables. Thus, in Fig. 5.6 both long bars are selected if no scene context is considered. If the bar context is considered the *seven-holed bar* is the only one selected.



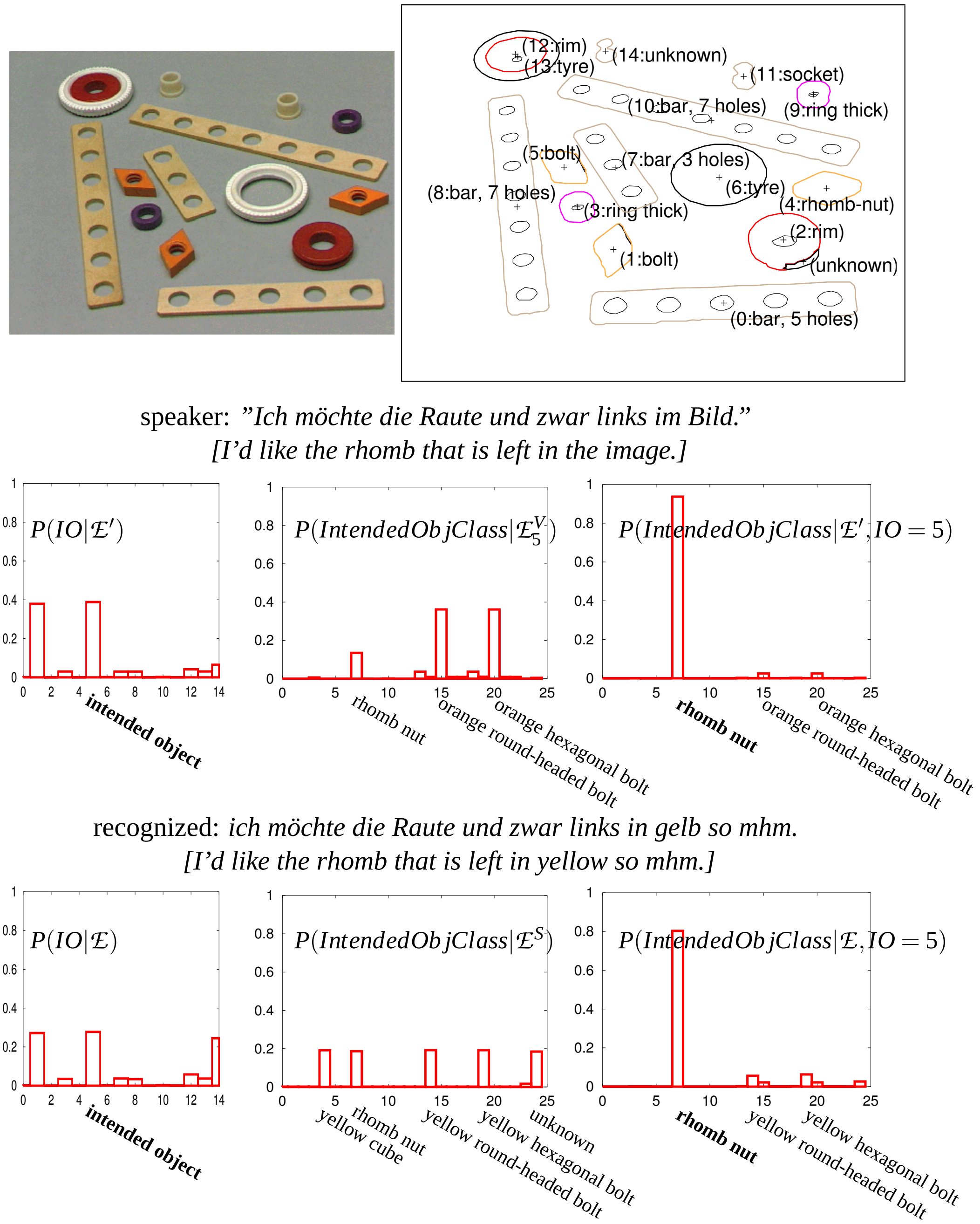
speaker: *"I möchte die helle lange Leiste." [I'd like the bright long bar.]*

Figure 5.6: Performance example 4: The speaker intends object 16. It is correctly recognized: $\mathcal{E}_{16}^V = \{wooden, 7\text{-}holed\text{-}bar\}$. The evidences extracted from speech are $\mathcal{E}^S = \{\text{hell, lang, Leiste}\}$. $\mathcal{E} = \bigcup_{i=0\ldots18} \mathcal{E}_i^V \cup \mathcal{E}^S$. 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.

Performance example 5 is shown in Fig. 5.7. Here the influence of speech recognition errors is examined. The speaker instructs the system by the utterance: *"I'd like the rhomb that is left in the image."* The speech recognition component decodes the word sequence: *"I'd like the rhomb that is left in yellow so mhm."*. Thus, the feature *yellow* is inserted into the verbal object description resulting in $\mathcal{E}^S = \{left, yellow, rhomb\}$.

The erroneous verbal object description results in a vague plausibility vector of the intended object class $P(IntendedObjClass|\mathcal{E}^S)$ (see lower second graph of Fig. 5.7). Further more, the intended object 5 has been misrecognized by the visual components. The speaker sees an *orange rhomb-nut*, the system detected an *orange bolt* (see second graph in the middle of Fig. 5.7).

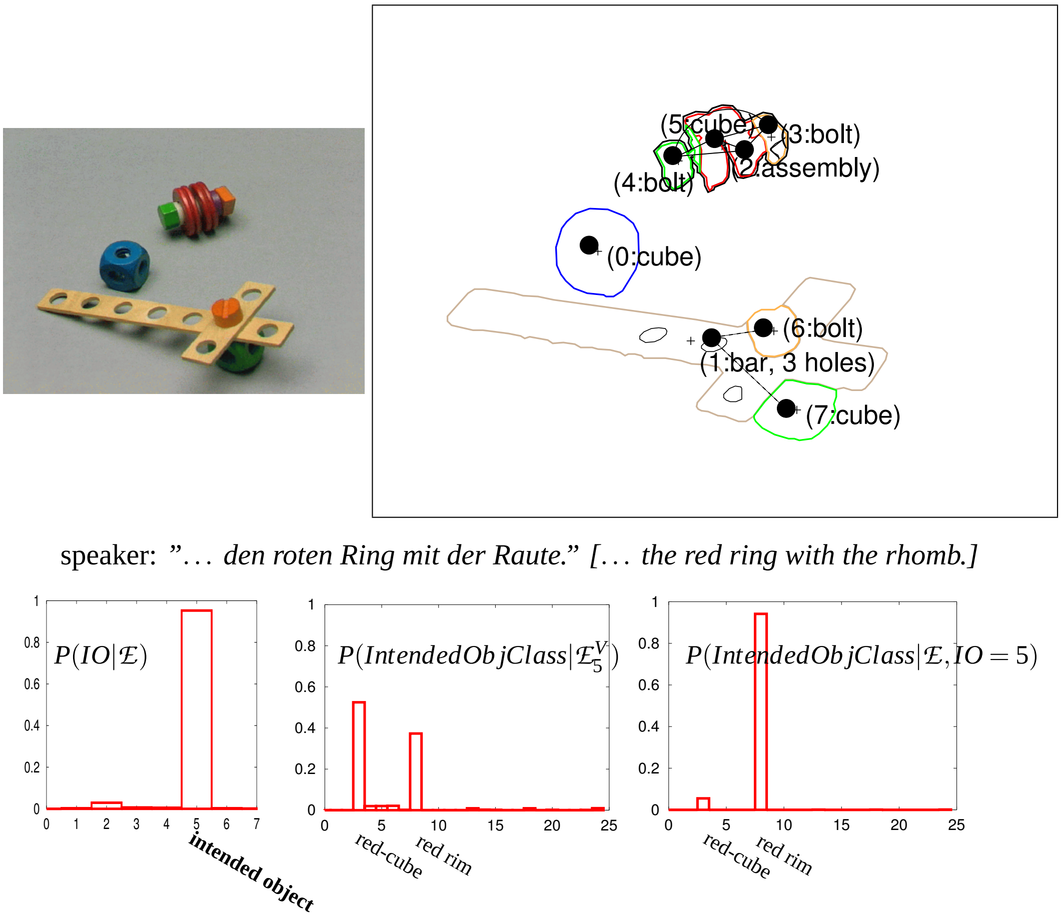
Nevertheless, the correct object is included in the systems answer consisting of the objects 1, 5, 14. The selection of the third object is caused by the speech recognition



speaker: *"Ich möchte die Raute und zwar links im Bild."*
[I'd like the rhomb that is left in the image.]

recognized: *ich möchte die Raute und zwar links in gelb so mhm.*
[I'd like the rhomb that is left in yellow so mhm.]

Figure 5.7: Performance example 5: The speaker refers to object 5. This rhomb-nut is incorrectly recognized: $\mathcal{E}_5^V = \{orange, bolt\}$. The evidences extracted from speech would have been $\mathcal{E}^{NL} = \{\text{links, Raute}\}$. $\mathcal{E}' = \bigcup_{i=0\ldots14} \mathcal{E}_i^V \cup \mathcal{E}^{NL}$. Instead the speech recognition errors yield a feature insertion: $\mathcal{E}^S = \{\text{links, gelb, Raute}\}$. $\mathcal{E} = \bigcup_{i=0\ldots14} \mathcal{E}_i^V \cup \mathcal{E}^S$. The system selects objects 1, 5 in the *NL* case, objects 1, 5, 14 with speech recognition errors. In any case for object 5 the correct class can be inferred.



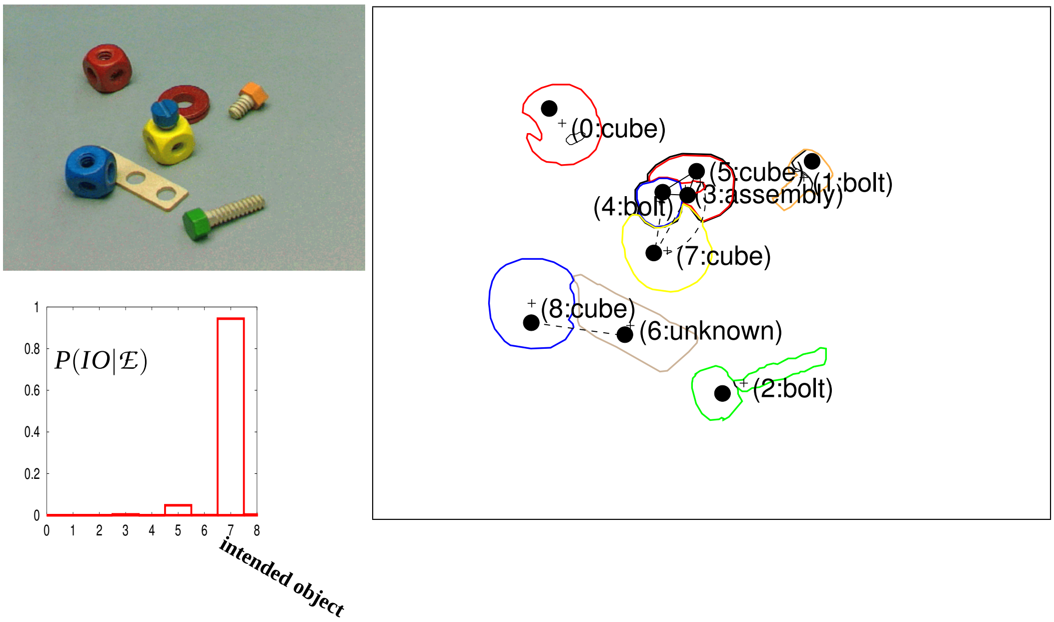
speaker: ”*. . . den roten Ring mit der Raute.*” *[. . . the red ring with the rhomb.]*

Figure 5.8: Performance example 6: The speaker refers to object 5 and reference object 3. The rim (obj. 5) and the rhomb-nut (obj. 3) are misrecognized: $\mathcal{E}_5^V = \{red, cube\}, \mathcal{E}_3^V = \{orange, bolt\}$. The mounting relation between them is correctly detected (solid lines). The dotted lines represent relations between touching object regions. The evidences extracted from speech are $\mathcal{E}_0^S = \{\text{rot, Ring}\}, \mathcal{E}_1^S = \{\text{mit, Raute}\}$. $\mathcal{E} = \bigcup_{i=0...7} \mathcal{E}_i^V \cup \mathcal{E}^S$. The system correctly selects object 5 with reference object 3. The class of object 5 *rim* is correctly inferred.

errors. Processing the correct NL input results in a selection of the objects 1, 5. Despite erroneous input in both channels the correct object class *rhomb-nut* is inferred for object 5 (see lower third graph in Fig. 5.7).

5.2.2 Interpretation of structural descriptions

Complex objects introduce additional aspects to the identification task (task 4, ***mp-struct***). On the one hand the visual analysis of complex objects is more difficult than that of isolated elementary objects. Assembled objects that have the same color result in segmentation failures. Occlusions cause erroneous classifications of object types. On the other



speaker: ”… *den gelben Würfel mit der Schraube.*”
[… *the yellow cube with the bolt.]*

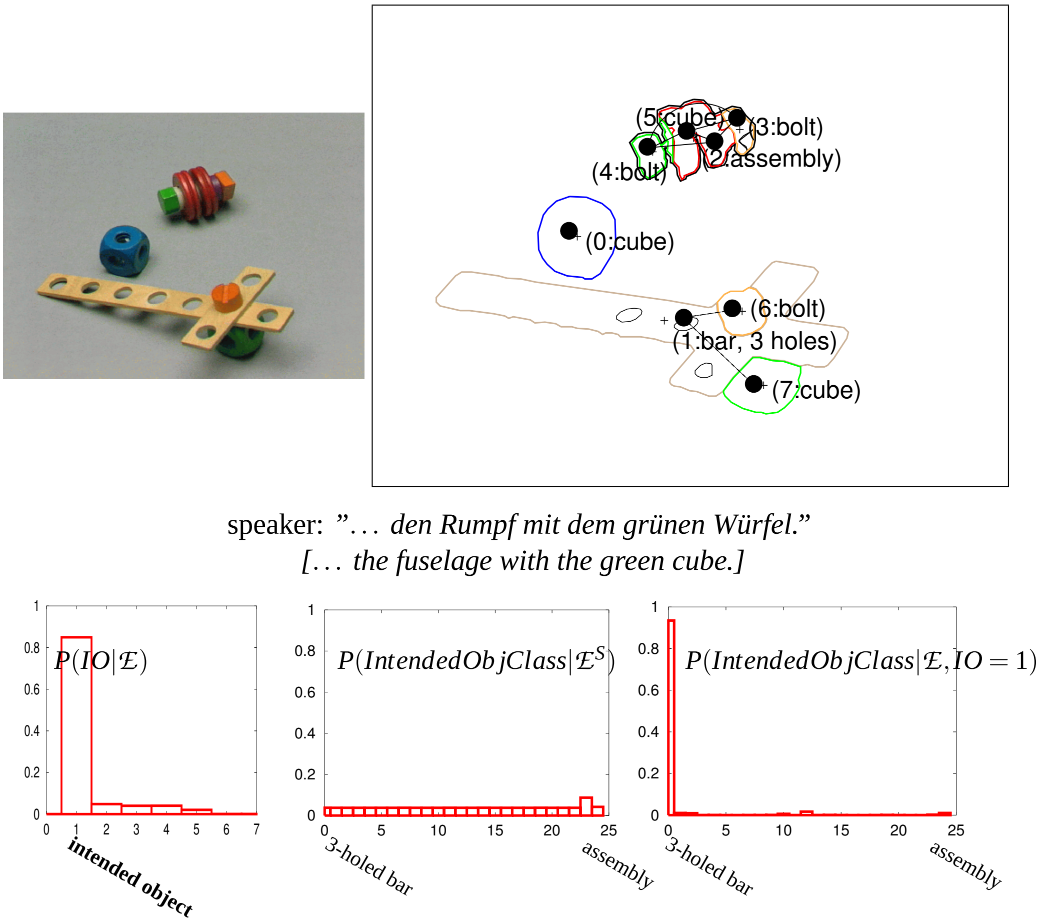
Figure 5.9: Performance example 7: The speaker refers to object 7 and reference object 4. Both object are correctly recognized: $\mathcal{E}_7^V = \{\text{yellow,cube}\}, \mathcal{E}_4^V = \{\text{blue, bolt}\}$. The mounting relation between them is not detected but the object regions touch each other (dotted edges). The evidences extracted from speech are $\mathcal{E}_0^S = \{\text{gelb, Würfel}\}, \mathcal{E}_1^S = \{\text{mit, Schraube}\}$. $\mathcal{E} = \bigcup_{i=0\ldots8} \mathcal{E}_i^V \cup \mathcal{E}_0^S \cup \mathcal{E}_1^S$. The system correctly selects object 7 with reference object 4.

hand verbal descriptions that specify structural relation between objects provide useful restrictions for object identification.

In performance example 6 presented in Fig. 5.8 three errors occurred in the visual analysis of two complex objects. Two *red rims* are classified as a *red-cube* due to segmentation failures, the *orange rhomb-nut* is classified as an *orange-bolt* due to occlusions, and the *3-holed-bar* and the *7-holed-bar* are incorrectly classified as a single *3-holed bar* due to segmentation failures. The speaker refers to one of the rims (obj. 5) of the first assembly: ”… *the red ring with the rhomb.*”

Although both specified objects have been misrecognized the intended object is correctly identified (Fig. 5.8, first graph). Considering both the visual and the verbal evidences the correct object type *red rim* is inferred (Fig. 5.8, third graph).

Performance example 7 demonstrates that verbal information can be used in order to establish hypotheses of un-recognized assembly structures (Fig. 5.9). The *red rim*



speaker: ”. . . *den Rumpf mit dem grünen Würfel.*”
[. . . the fuselage with the green cube.]

Figure 5.10: Performance example 8: The speaker refers to the complex object consisting of the objects 1,6,7 that is not detected. Object 1 is misrecognized: $\mathcal{E}_1^V = \{wooden, 3\text{-}holed\text{-}bar\}$. Object 7 is correctly recognized: $\mathcal{E}_7^V = \{green, cube\}$. The word *fuselage* is an unknown object name: $\mathcal{E}_0^S = \{\text{Objekt}\}$. The other speech evidences are $\mathcal{E}_1^S = \{\text{mit, grün, Würfel}\}$. Instead of the whole complex object the system selects object 1 with reference object 7. The unknown object name *fuselage* is interpreted as another name for the *3-holed-bar*.

(obj. 5) has been misrecognized as a *red-cube*. Consequently, the *orange-bolt* and the misrecognized *red-cube* are hypothesized as a complex object leaving the *yellow-cube* as an isolated elementary object. The speaker introduces a structural relation between the *yellow-cube* and the *orange-bolt* by the instruction: ”. . . *the yellow cube with the bolt.*”

The yellow cube is correctly identified (graph in Fig. 5.9) and, consequently, can be hypothesized as the part of a complex object consisting of the yellow cube and the orange bolt.

5.2.3 Unknown object names

Complex objects are difficult to name. The speaker can use the abstract name *assembly* or describe the structure of the complex object. Most frequently complex objects have a functional role in the construction process. For example they are the *rotor, cockpit, fuselage, motor, wing*, or the *tail* of a *toy-airplane*. However, the system cannot be sure that in any case a complex object is denoted by these names. Even elementary object can have such a role (task 5, ***mp-name***).

In performance example 8 the unknown object name *fuselage* is linked to the elementary object 1, the *3-holed-bar* (Fig. 5.10). The speaker describes the intended object by the utterance: ”*... the fuselage with the green cube.*”.

The structural relation is mapped by the system to the objects 1 and 7 despite the mounting relation was not detected. Due to the misrecognized bars the mounting ports of the objects did not fit. However, the system detected that the two object regions touch each other (dotted lines in Fig. 5.10). Although the errors introduced by the vision components could not be totally corrected the system answer is partially correct.

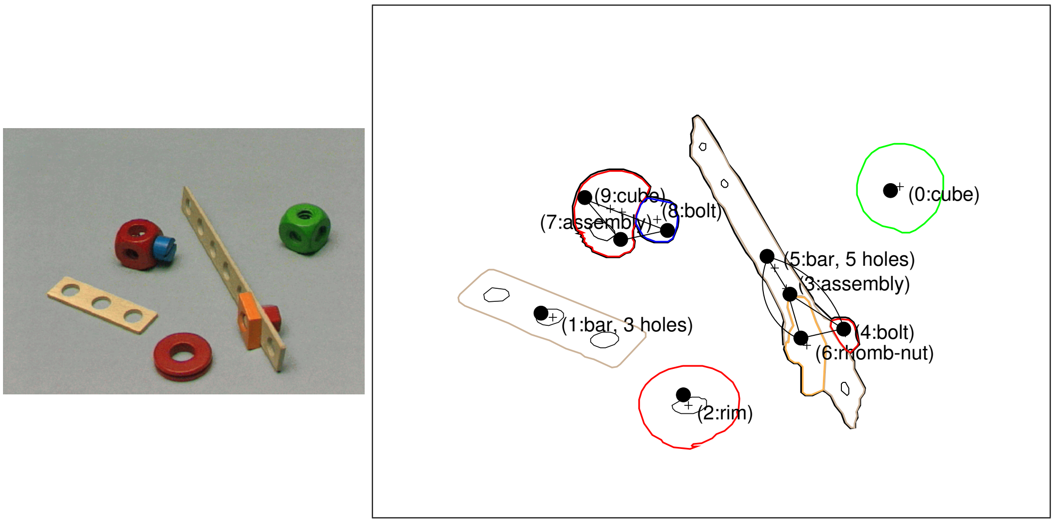
Performance example 9 demonstrates how unknown object names of complex objects can be learned by the system (Fig. 5.11). The speaker describes the complex objects by the utterances: ”*... the motor behind the bar*” and ”*... the fuselage to the right of the motor*”.

Both assemblies have been correctly recognized. The spatial relations constrain the system selection so that the correct visual objects are linked to the unknown object names. If the result of the first utterance is stored by the system it could be used in the interpretation of the second utterance. By this means, the system is able to increase its competence. The treatment of unknown object names as presented in this thesis provides the basis for such a learning strategy.

5.2.4 Disambiguating alternative interpretations of an utterance

Some utterances cannot be interpreted in an unique way. They are syntactically or semantically ambiguous (task 1, ***mp-interp***). In the following performance examples prepositional attachments and extrapositions will be discussed. In both cases the syntactical structure of the sentence is ambiguous and, consequently, leads to different verbal object descriptions. In order to figure out the intended object all possible interpretations have to be considered. The interpretation that reveals the most probable selection of an object is assumed to be the intended meaning.

Let $\mathcal{E}_A^S$ represent the speech evidences of a first interpretation A including a reference object RO, let $\mathcal{E}_B^S$ represent the speech evidences of an alternative interpretation B including no reference object. Let $\mathcal{E}_{0\ldots n}^V$ be the visual evidences of the n scene objects. Then, the Bayesian network has to be evaluated for the evidential sets $\mathcal{E}_A = \bigcup_{i=0\ldots n} \mathcal{E}_i^V \cup \mathcal{E}_A^S$



1. speaker: ”... *den Motor hinter der Leiste.” [the motor behind the bar.]*

2. speaker: ”... *den Rumpf rechts von dem Motor.” [the fuselage to the right of the motor.]*

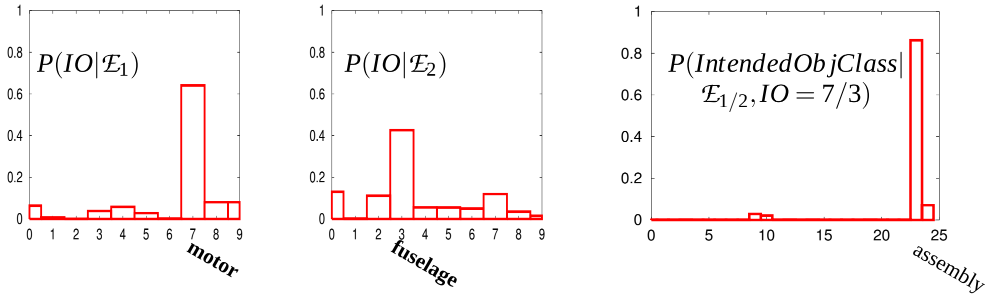


Figure 5.11: Performance example 9: The speaker refers to the complex objects 7 (fst. utterance) and 3 (snd. utterance). Both assemblies are correctly recognized: $\mathcal{E}^V_{3/7} = \{assembly\}$. The evidences extracted from speech are $\mathcal{E}^{S_1}_0 = \{\text{Objekt}\}, \mathcal{E}^{S_1}_1 = \{\text{hinter, Leiste}\}$ and $\mathcal{E}^{S_2}_0 = \{\text{Objekt}\}, \mathcal{E}^{S_2}_1 = \{\text{rechts-von, Objekt}\}$. $\mathcal{E}_1 = \bigcup_{i=0...9} \mathcal{E}^V_i \cup \mathcal{E}^{S_1}$. $\mathcal{E}_2 = \bigcup_{i=0...9} \mathcal{E}^V_i \cup \mathcal{E}^{S_2}$. The system correctly selects the objects 7 with reference object 1 and object 3 with reference object 7.

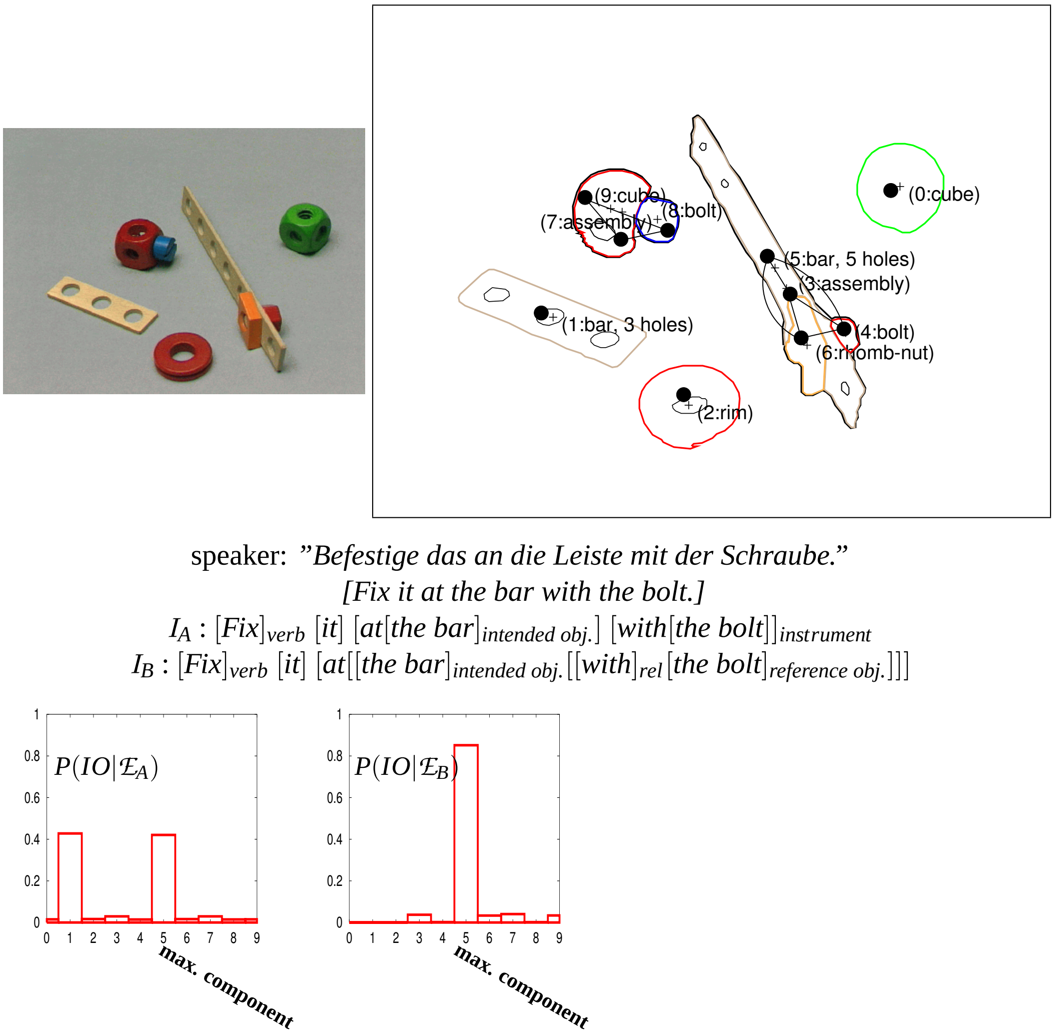
and $\mathcal{E}_B = \bigcup_{i=0...n} \mathcal{E}^V_i \cup \mathcal{E}^S_B$:

$$(io^*_A, ro^*_A) = \underset{io,ro}{\operatorname{argmax}} P(IO = io, RO = ro|\mathcal{E}_A), \quad \vec{\eta}^A = \beta \max_{ro} P(IO, RO = ro|\mathcal{E}_A)$$

$$(io^*_B) = \underset{io}{\operatorname{argmax}} P(IO = io|\mathcal{E}_B), \quad \vec{\eta}^B = P(IO|\mathcal{E}_B)$$

$$(io^*, ro^*, \mathcal{E}) = \begin{cases} (io^*_A, ro^*_A, \mathcal{E}_A), & \text{if } (\max_i \eta^A_i) \geq (\max_j \eta^B_j) \\ (io^*_B, \perp, \mathcal{E}_B), & \text{otherwise} \end{cases}$$

By this means, alternative interpretations of a spoken utterance can be disambiguated.



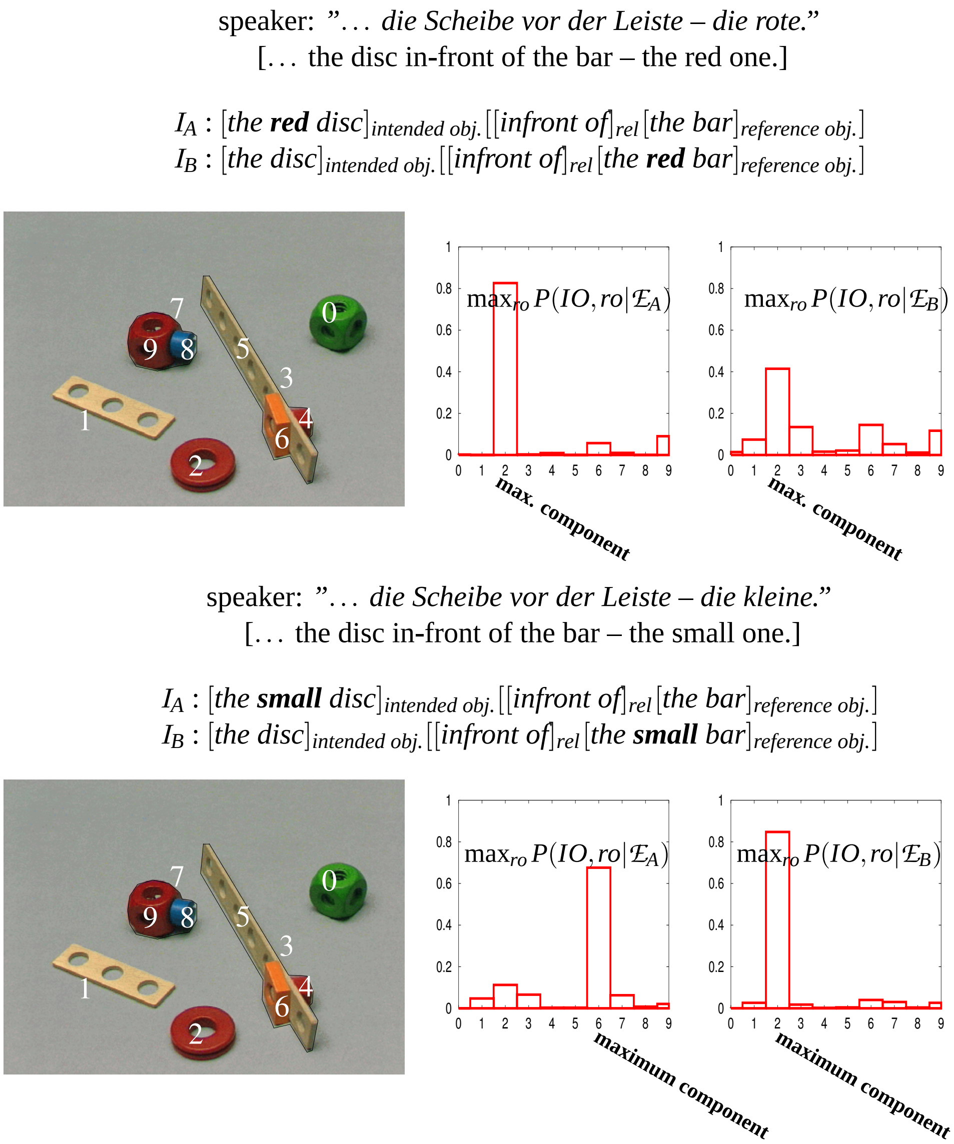
speaker: *"Befestige das an die Leiste mit der Schraube."*
[Fix it at the bar with the bolt.]
I_A : $[Fix]_{verb}$ $[it]$ $[at[the\ bar]_{intended\ obj.}]$ $[with[the\ bolt]]_{instrument}$
I_B : $[Fix]_{verb}$ $[it]$ $[at[[the\ bar]_{intended\ obj.}[[with]_{rel}[the\ bolt]_{reference\ obj.}]]]$

Figure 5.12: Performance example 10: The speaker refers to object 5 using a structural description. The utterance is ambiguous because the prepositional attachment can also be interpreted as an instrument. The first interpretation I_A yields the evidence $\mathcal{E}_A^S = \{\text{Leiste, mit, Schraube}\}$. $\mathcal{E}_A = \bigcup_{i=0\ldots9} \mathcal{E}_i^V \cup \mathcal{E}_A^S$. The second interpretation I_B yields the evidence $\mathcal{E}_A^S = \{\text{Leiste}\}$. $\mathcal{E}_B = \bigcup_{i=0\ldots9} \mathcal{E}_i^V \cup \mathcal{E}_B^S$. The system decides for interpretation A and selects object 5.

Performance example 10 demonstrates the disambiguation of a prepositional attachment (Fig. 5.12). The speaker instructs the system by the utterance: *"Fix it at the bar with the bolt."*

The prepositional phrase *[with the bolt]* can either be attached to the verb *[fix]* or to the noun phrase *[the bar]*. Thus, two different verbal object descriptions have to be considered: $\mathcal{E}_A^S = \{bar\}, \mathcal{E}_B^S = \{bar, with, bolt\}$.

In the visual context (see Fig. 5.12) an isolated *3-holed-bar* (obj. 1) and a *7-holed-bar*



speaker: ”*. . . die Scheibe vor der Leiste – die rote.*”
[. . . the disc in-front of the bar – the red one.]

I_A : $[$*the* ***red*** *disc*$]_{\textit{intended obj.}}[[$*infront of*$]_{\textit{rel}}[$*the bar*$]_{\textit{reference obj.}}]$
I_B : $[$*the disc*$]_{\textit{intended obj.}}[[$*infront of*$]_{\textit{rel}}[$*the* ***red*** *bar*$]_{\textit{reference obj.}}]$

speaker: ”*. . . die Scheibe vor der Leiste – die kleine.*”
[. . . the disc in-front of the bar – the small one.]

I_A : $[$*the* ***small*** *disc*$]_{\textit{intended obj.}}[[$*infront of*$]_{\textit{rel}}[$*the bar*$]_{\textit{reference obj.}}]$
I_B : $[$*the disc*$]_{\textit{intended obj.}}[[$*infront of*$]_{\textit{rel}}[$*the* ***small*** *bar*$]_{\textit{reference obj.}}]$

Figure 5.13: Performance example 11: The visual context is that of Fig. 5.12. The speaker refers to the red rim (obj. 2) in both utterances. The utterances are syntactically ambiguous because the extrapositions *’the red one’* and *’the small one’* may either extend the intended object description (I_A) or the reference object description (I_B). The system selects interpretation A for the first and interpretation B for the second utterance.

(obj. 5) that is part of a complex object are detected. Therefore, the verbal description *[the bar]* is referentially ambiguous (first graph in Fig. 5.12). The alternative description *[the bar with the bolt]* can be precisely decoded to refer to the *5-holed-bar* (obj. 5, see second graph in Fig. 5.12). Therefore, interpretation B is selected.

Performance example 11 shows the disambiguation of an extrapositions (Fig. 5.13). The system results for both utterances have been calculated. In the first one ”… *the disc in-front of the bar – the red one”* the extraposition *’the red one’* shall extend the description of the intended object *’disc’*. In the second utterance ”… *the disc in-front of the bar – the small one”* the extraposition *’the small one’* shall extend that of the reference object *’bar’*. However, both cases are syntactically ambiguous because the extrapositions may extend both noun phrases in the sentences.

The visual context is the same as in the previous performance example (Fig. 5.12). The visual scene includes two bars, a short isolated one (obj. 1) and a long one (obj. 5) that is part of an assembly. The name *disc* may fit on two possible objects, the *red rim* (obj. 2) and with less probability the *orange rhomb-nut* (obj. 6).

In both alternative interpretations of the first utterance, ”… *the disc in-front of the bar – the red one”*, the system correctly selects the *red rim* (obj. 2) and reference object 1 (*wooden, 3-holed-bar*) but the second interpretation includes contradictory evidences (*red, bar*). Thus, the first interpretation is selected (see upper graphs in Fig. 5.13).

For the second utterance, ”… *the disc in-front of the bar – the small one.*”, the alternative interpretations cause the selection of different object pairs. In the first case the *small disc* is linked to the *orange rhomb-nut* (obj. 6) and the *bar* is linked to the *7-holed-bar* (obj. 5). In the second case the system correctly selects object 2 (*red rim*) and reference object 1 (*3-holed-bar*). This time the second interpretation better explains the detected evidences (see lower graphs in Fig. 5.13).

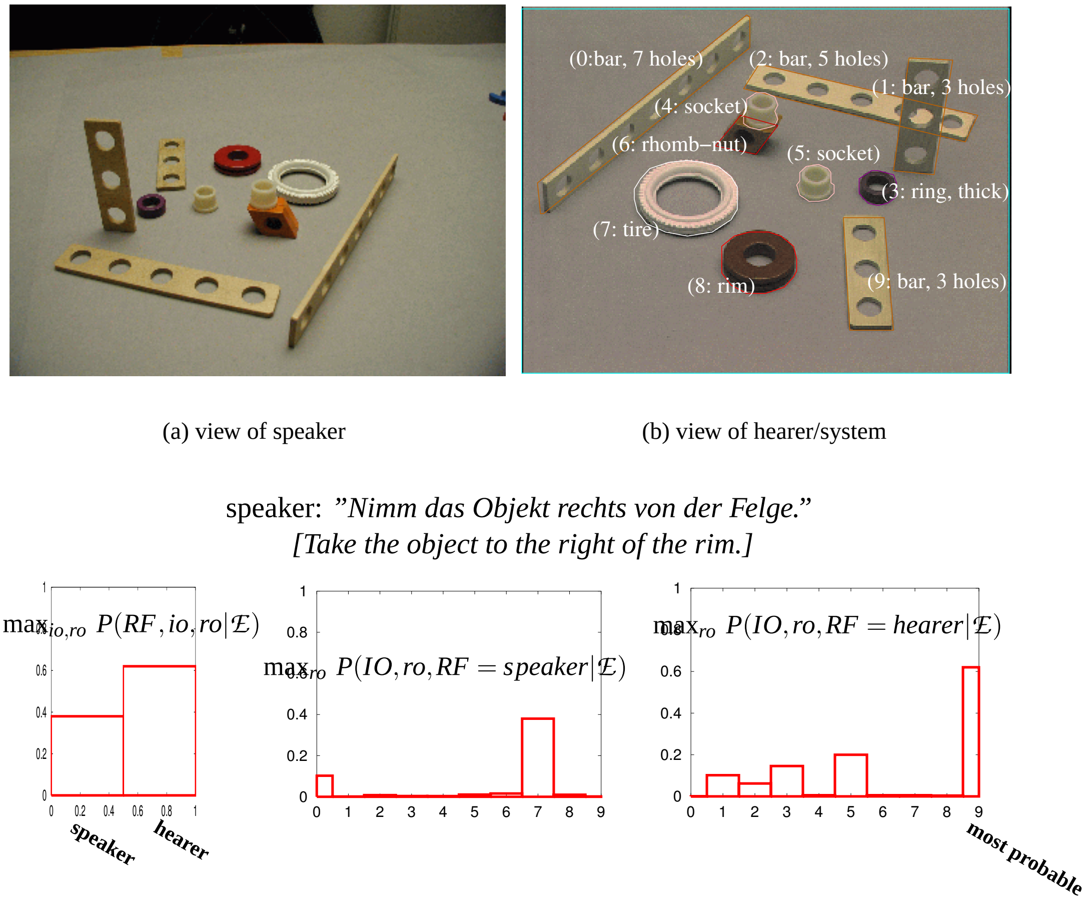
5.2.5 Disambiguating the selected reference frame

In the previous chapter it was argued that a more detailed Bayesian modeling of the spatial model can be used in order to estimate the selected reference frame of the speaker. In this case the random variable *RF* that denotes the reference frame is added to the query variables:

$$(rf^*, io^*, ro^*) = \underset{rf,io,ro}{\operatorname{argmax}} P(RF = rf, IO = io, RO = ro|\mathcal{E}) \tag{5.5}$$

In the following performance example two possible reference frames are considered. It is assumed that the speaker and the cameras of the system, i.e. the hearer, face each other. Consequently, the *hearer-centered* reference frame is rotated with regard to the *speaker-centered* by 180 degree. The speaker instructs the system by the sentence: ”*Take the object to the right of the rim.*”

It includes the implicit selection of a reference frame. The scene context and evaluation results are given in Fig. 5.14. The *3-holed-bar* is directly located at the horizontal axes through the rim. This is the prototypical direction for the projective relation *right-of* with regard to the *hearer-centered* reference frame. Therefore, it is selected as the most probable hypothesis. From the identification result of the query the implicit use of the reference frame can be inferred, here the *hearer-centered* reference frame.



(a) view of speaker

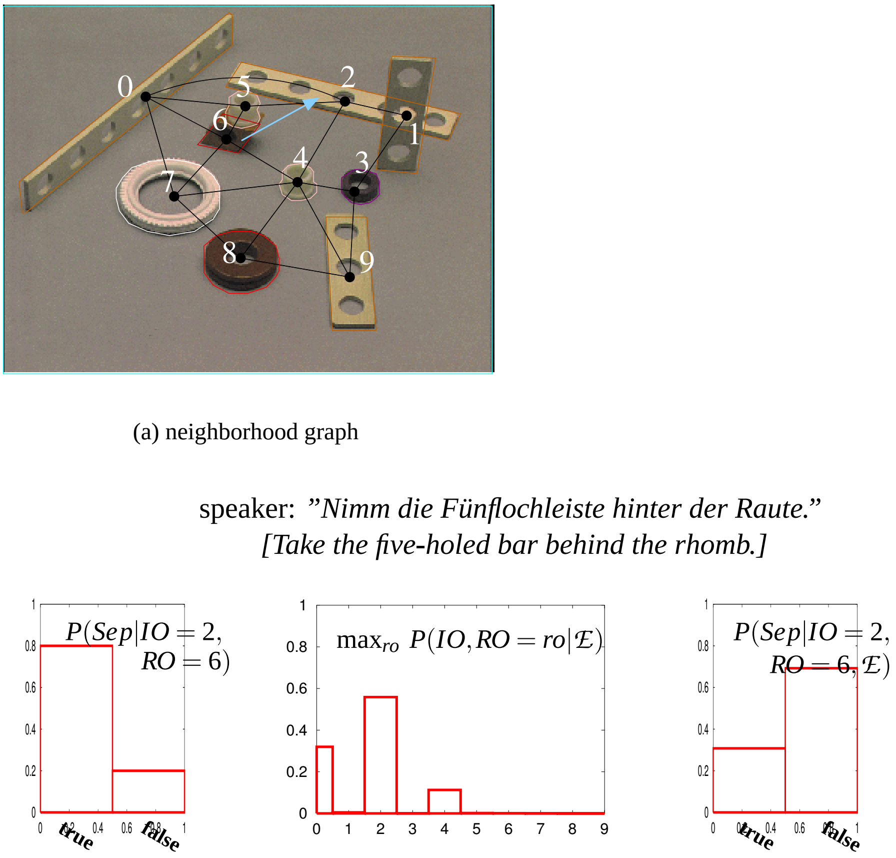
(b) view of hearer/system

speaker: *"Nimm das Objekt rechts von der Felge."*
[Take the object to the right of the rim.]

Figure 5.14: Performance example 12: The speaker refers to object 9. He has selected a hearer-centered reference frame for the mentioned projective spatial relation. The first graph shows the normalized result for the most probable hypotheses with speaker-centered and hearer-centered reference frames. The next two graphs show the plausibility for all possibly intended objects with regard to both reference frames. The most probable hypothesis is the selection of the triple $IO = 9, RO = 8, RF =$ *hearer-centered*.

5.2.6 Detection of neighborhood relations

Another aspect of the more detailed spatial network is the explicit modeling of the neighborhood relations. The computational model (see Sec. 4.4.3) is used to calculated the a priori probabilities $P(Sep|IO, RO)$. The random variable *Sep* has two states $\{$*true, false*$\}$ that denote if two objects are separated by other objects, i.e. are *not neighboring*, or not, i.e. are *neighboring*. The computational model provides an expectation about the separation and neighborhood of objects. This expectation will not always be true. A speaker may select a reference object in a verbal object description that is *separated* from the intended object due to the threshold criterion of the computational model. Applying the assumption that a speaker will always use neighboring objects in such a spatial description



(a) neighborhood graph

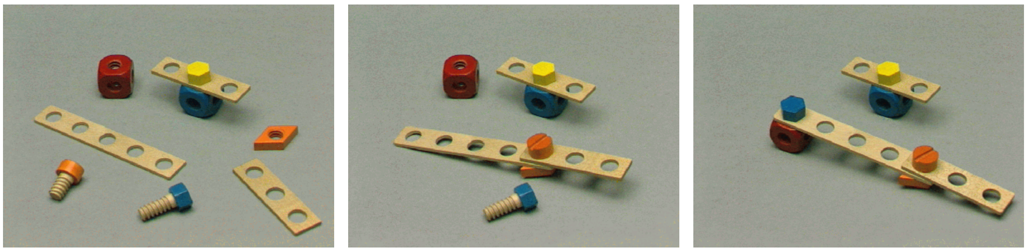
speaker: *"Nimm die Fünflochleiste hinter der Raute."*
[Take the five-holed bar behind the rhomb.]

Figure 5.15: Performance example 13: The speaker refers to object 2 with reference object 6. In the neighborhood graph there exists no edge between the corresponding nodes because the object regions are separated by object 5. Thus, in the first graph the separation variable is expected to be *true*. Nevertheless the system correctly selects the object 2 and reference object 6 for the utterance. The third graph shows the updated expectation for the state of the separation variable. The most probable state is *false*.

this can be interpreted as an additional evidence for neighborhood and non-separation.

In Fig. 5.15 such an example is given. The speaker intends to describe the *5-holed-bar* (obj. 2) and uses the *rhomb-nut* (obj. 6) as a reference object: *"take the five-holed bar behind the rhomb."*

In this example the visual data is given by the hand labeled image, i.e. segmentation and classification errors are excluded. The computational model detects a separation of these two objects that is caused by the *socket* on the *rhomb-nut*. However, this separation is only a perspective artifact. In three dimensions no object is placed between them. The a priori probability for the object separation, that is given by the computational model, is shown in the first graph of Fig. 5.15. Nevertheless, the correct intended and reference



(a) ”First we construct the fuselage . . . ”

(b) ”. . . fix the cube at the five-holed bar.”

(c) ”Now, we have added the engine.”

(d) ”The other propeller engine block will be used later.”

Figure 5.16: An example for a construction dialog. The unknown names *fuselage, engine, and propeller* denote subassemblies.

objects are obtained (see second graph of Fig. 5.15). Considering these evidences the a posteriori probability for an object separation can be calculated (third graph in Fig. 5.15). It follows from the above that objects 2 and 6 are *not separated* and, consequently, are *neighboring*.

5.3 Further Learning Capabilities

The previous sections describe how elementary or complex objects can be linked to verbal descriptions and especially to *unkown names* (Sec. 5.2.3). However, the problem of learning such denotations becomes even more complicated if subparts of a complex object are named by a speaker. This may frequently happen during a construction dialog. The system has to find out which subset of elementary objects is denoted. In most cases this task cannot be solved considering a single situation. Thus, ***cross-situation learning*** strategies (cf. Sec. 2.5.3) may be employed that are applied during each dialog step.

The idea may be clarified by an example (Fig. 5.16). The speaker intends to construct an airplane and starts the construction dialog by the following utterance:

”First we construct the fuselage . . . ”

After several construction steps, the next unknown name is introduced:

”. . . fix the cube at the five-holed bar. Now, we have added the engine.”

The system does not have any information when the construction of the fuselage ended and the construction of the engine began. However, the system can infer that the elementary object used lastly must be an element of the engine. Then the speaker refers to the other complex object in the scene consisting of a *bolt*, a *3-holed-bar*, and a *cube*:

”The other propeller engine block will be used later.”

From these information the meaning of propeller, engine, and fuselage can be directly inferred using the inference rules proposed by Siskind [Sis96] (cf. Sec. 2.5.3).

The meaning of an unkown name w is represented as three different sets[1]:

1. $\mathcal{P}(w)$ is the set of object types that are possibly involved in the meaning of w.
2. $\mathcal{N}(w)$ is the set of object types that are necessarily involved in the meaning of w.
3. $\mathcal{D}(w)$ is the set of possible assembly structures that denote the meaning of w.

In the previous examples the inference rules are applied as follows[2]:

1. *"First we construct the fuselage . . . "*

   Initially, the necessary set is empty and the possible set contains all possible object types:

	$\mathcal{N}$	$\mathcal{P}$
fuselage	{}	{*3-holed-bar, 5-holed-bar, 7-holed-bar, cube, rim, tire*}

2. *". . . ... fix the cube at the five-holed bar. Now, we have added the engine."*

   The construction of the fuselage and the engine is finished. Rule 2 states that those words can be eliminated from the possible sets that are not included in one of the considered meanings. Note that the fuselage has been finished before the last instruction was executed:

	$\mathcal{N}$	$\mathcal{P}$
fuselage	{}	{*3-holed-bar, 5-holed-bar*}
engine	{}	{*3-holed-bar, 5-holed-bar, cube*}

   Then, we apply rule 3 adding those members of $\mathcal{P}(w)$ to $\mathcal{N}(w)$ that are not a member of any other possible set $\mathcal{P}(w')$:

	$\mathcal{N}$	$\mathcal{P}$
fuselage	{}	{*3-holed-bar, 5-holed-bar*}
engine	{*cube*}	{*3-holed-bar, 5-holed-bar, cube*}

3. *"The other propeller engine block will be used later."*

   Now the representation of the other complex object is used to constrain the meaning of the names further. Rule 2 eliminates the *5-holed-bar* from $\mathcal{P}$(engine):

	$\mathcal{N}$	$\mathcal{P}$
fuselage	{}	{*3-holed-bar, 5-holed-bar*}
engine	{*cube*}	{*3-holed-bar, cube*}
propeller	{}	{*3-holed-bar, cube*}

   Rule 4 states that each symbol that appears only once in every remaining utterance

[1] In the sets only bars, cubes, rims, and tires are considered because these elementary objects are the backbone of any complex object.

[2] In order to simplify the example only the possible and necessary sets are shown

meaning can be eliminated from the possible set if it is member of the necessary set of another word. Thus, the *cube* can be eliminated from $\mathcal{P}(\text{propeller})$. Then, an additional rule must be applied that takes account of the fact that any meaning of a name must consist of one elementary object at minumum. Consequently, the *3-holed-bar* can be added to $\mathcal{N}(\text{propeller})$ ruling out this object for the $\mathcal{P}(\text{engine})$ set (rule 4). Now, rule 3 can be applied to the previous denotation adding the *3-holed-bar* and the *5-holed-bar* to the necessary set of the fuselage $\mathcal{N}(\text{fuselage})$:

	$\mathcal{N}$	$\mathcal{P}$
fuselage	{*3-holed-bar, 5-holed-bar*}	{*3-holed-bar, 5-holed-bar*}
engine	{*cube*}	{*cube*}
propeller	{*3-holed-bar*}	{*3-holed-bar*}

In the last step, the elementary object sets of the meaning representation of the previously unknown names *fuselage*, *engine*, and *propeller* have converged. The set $\mathcal{D}(w)$ denote the possible assembly structures that can be built from the elementary objects of the possible set. Subsequently, this set can be restricted by considering the extracted assembly structures of the complex objects that were denoted, and thereby the meaning of the names can be extracted.

The presented technique has not been implemented so far and goes beyond the current performance of the system. However, the idea seems to be promising to be integrated into the system. As a consequence, an identification of complex objects will result in a structural comparison of assemblies. This issue is only primarily treated by this thesis.

Chapter 6

Results

In the last chapter the functionality of the integration component has been demonstrated by several performance examples. However, the robustness of the system must be shown by a quantitative measurement of identification results.

6.1 Test Sets

The evaluation experiments have been performed on data collected in the experiments 5 (Select-Obj) and 6 (Select-Rel) (see Sec. 4.2). In both experiments subjects verbally described marked objects in table scenes that were presented on a computer screen.

The data sets were used on different processing levels. Thus, the integration component can be tested under different error conditions and the influence of different kinds of uncertainty can be measured:

- speech data:

  1. On the most abstract level the object features mentioned by the speaker were transcribed in a form that can be directly processed by the integration component. This verbal data is called ***FEATURE***.
  2. The interface between the speech recognition and speech understanding component consists of the most probably uttered word sequence. These recognized word sequences were transcribed into feature structures that can be directly processed by the integration component. This input data is called ***FEAT_SPEECH***. The comparison with the ***FEATURE*** set measures the influence of speech recognition errors on identification results assuming a perfect understanding component.

- image data:

  1. The most abstract visual data is given by hand labeled object regions on the image plane. Segmentation and classification errors are thereby excluded. Interpretation errors can still be caused by perspective artifacts from the 2-d

camera view, e.g. illegal object separations. This set is called ***OBJECTS***. The processing is directly started with the calculation of the spatial models of the integration component.

2. The second visual level consists of YUV-images taken by the camera. Therefore, processing starts with segmentation and object recognition. This set is called ***IMAGES***.

The evaluation of the speech and object recognition components are presented in detail in Sec. 4.3.3. However, the word, feature, and detection accuracies will be repeated in the following subsections.

6.2 Classification of System Answers

Given an instruction by the speaker and an image from the table scene the system identifies the objects that were hypothetically denoted by the instruction. The correct object is given by the marked one. However, this does not imply that a perfect system would identify the marked object in all identification tasks. Some tasks are only solved partly because the object was not precisely specified. Nevertheless, the marked object is the reference data for the evaluation of the system answer. The following overlapping classes are distinguished:

- ***precise***: the marked object is the only one the system selected.

- ***included***: the system may have selected more objects besides the marked one. But all selected objects have the correct unique object type. Note that ***precise*** is a subset of ***included***.

- ***additional***: the marked object is member of the selected set of objects. But some selected objects have a different unique object type than the marked one.

- ***correct***: the marked object is a member of the selected subset. Note that ***correct*** is the union of ***included***, and ***additional***.

- ***false***: the system has selected some objects, but the marked one is not a member of the subset.

- ***nothing***: the system has rejected the instruction because the system did not find an appropriate object.

The precise class is only relevant if spatial relations or localization attributes have been mentioned by the speaker. Therefore, it is especially relevant in the *Select-Rel* set when the subjects were explicitly told to use a spatial relation.

The most important error rates are those counted by the ***correct*** and ***additional*** categories. As long as the number of additional objects is small the system answers are acceptable for the user because he can select the intended object in the next dialog step.

6.3 Results on the *Select-Obj* test set

The *Select-Obj* test set consists of 453 utterances. The 10 different speakers describe a marked object on a computer screen by type, color, shape, size, localization attributes, or spatial relations (see Sec. 4.2 for a detailed description of the experiment).

The word accuracy (WA) of the speech recognition results on this test set is 68.2%, the feature accuracy (FA) is 85.0%. For 6 out of 453 utterances speech recognition errors caused a complete misinterpretation of the verbal object description. Either the object features were totally lost due to word deletions or substitution, or the intended object was interpreted as a reference object because of an insertion of a spatial relation. These utterance were ignored in the FEAT_SPEECH tests.

The scenes that were used as the visual context contain only elementary objects, no complex objects. In some of the scenes out-of-domain objects, like a screw-wrench, a toy-bus, or a cloth, were placed on the table. The detection accuracy (DA) of the objects in the table scenes was 74.6%. The recognition errors include 17.6% false type classifications and 4.9% false color classifications.

Before the identification rates of the system will be discussed, the expected system behavior should be clarified:

- If the speaker verbally refers to the marked objects, this should be included in the answer of the system. Consequently, we expect a ***correct*** rate of near 100% if no recognition errors occurred.

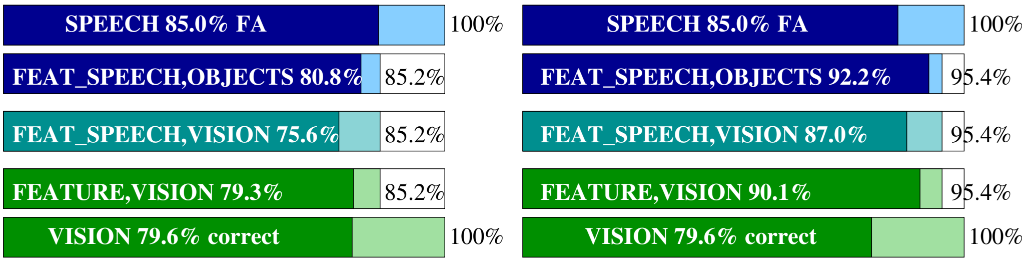
- In this test set, most speakers describe the type of the marked object but do not locate them precisely. Consequently, the ***included*** rate should be very high. However, some utterances even do not exactly describe the type of the object, e.g. *”und jetzt die Lochleiste.”* *[And now the holed bar.]* if there are three-holed and five-holed bars in the scene.

- Because the speakers often do not describe the location of the marked object the ***precise*** rate will be lower. In many utterances the speaker *intends* to specify a subset of objects, e.g. *”Alle gelben eckigen Muttern.”* *[All yellow angular nuts.]*

- If some features that were mentioned by the speaker are misrecognized or not recognized at all, *the identification rates will decrease*. 79% of the verbal object descriptions have been correctly recognized (cf. Sec 4.3.3). Therefore, the decrease of the identification rates should be 21% at maximum. The counting of complete object descriptions might not be a good measure for evaluating the impact of speech recognition errors because the rate does not take into account verbal object descriptions that are partially correct. Therefore, the feature accuracy will be treated as a measure for the average impact of speech recognition errors. The feature accuracy of 85% equals a feature error rate of 15%. The same error rate would result if 15% of the speakers’ descriptions were completely misrecognized. Consequently, we expect an average decrease of 15% of the identification rates.

Select-Obj		#utt	precise	included	additional	correct	false
FEATURE	OBJECTS	453	52.6%	**85.2%**	10.2%	**95.4%**	4.6%
FEAT_SPEECH	OBJECTS	447	51.2%	**80.8%**	11.4%	**92.2%**	7.8%
FEATURE	VISION	453	50.5%	**79.3%**	10.8%	**90.1%**	9.9%
FEAT_SPEECH	VISION	447	48.5%	**75.6%**	11.4%	**87.0%**	13.0%

(a) identification rates



(b) included

(c) correct

Table 6.1: Identification results for the *Select-Obj* set. For this test set the most relevant rates are those of the *included* and *correct* categories. The subfigures (b) and (c) visualize the impact of erroneous input data. The feature accuracy (top beam) is measured on the FEAT_SPEECH subset. The *correct* object recognition rate (bottom beam) is measured on those visual objects that should be identified. The light colors denote the base-line results from the FEATURE-OBJECTS data. The dark blue color indicates the impact of speech recognition errors, the dark green color indicates the impact of object recognition errors. The dark cyan beam visualizes the identification rates if both impacts are considered.

- If some of the *marked objects* have been misrecognized, *the identification rate will decrease*. 79.6% of the marked objects are correctly recognized. Therefore, the expected decrease of the identification rate is about 20%. The rate may additionally be affected by other misrecognized objects (DA 74.6%) because these may be used as a reference object or are also selected due to a misrecognized type or color.

The identification rates of the integration component are presented in Tab. 6.1. If speech and object recognition errors are excluded (FEATURE-OBJECTS) in only 4.6% of the utterances the marked object is not selected. A typical example for a false system answer is the following situation:

> *The speaker refers to a marked orange rhomb-nut by the utterance "Die mittlere rote Scheibe." [The red disc in the middle]. The system detects the orange rhomb-nut in the middle and a red rim on the right side. The red rim is selected.*

The low *false* identification rate of 4.6% shows that the baufix®domain is adequately modeled by the integration component.

The FEATURE-OBJECTS identification rates define the ***base-line results*** for the evaluation of system answers with erroneous input data. In Tab. 6.1(b) and 6.1(c) these are plotted with light colors. The *included* and *correct* identification rates with erroneous input data are plotted with dark colors in order to visualize the magnitude of decrease:

- The error rates of the input data are much higher than the decrease of the identification rates with regard to the base-line results. Even the decrease of the FEAT_SPEECH-VISION identification rates (11.3% for *included*, 8.8% for *correct*) is smaller than both input error rates (15.0% for SPEECH, 20.4% for VISION).

- The *correct* identification rates are more stable than the *included* rates. This should be expected for a robust system behavior. Errors do not cause completely wrong system answers but the quality of the answers decrease.

- Both noisy input channels decrease the identification rates by a similar magnitude. In the combined case the impacts on the identification rate are nearly additive. This indicates that both influences may be independent.

6.4 Results on the *Select-Rel* test set

In the *Select-Rel* experimental setting the subjects were instructed to use spatial relations in order to describe the marked object. Therefore, the collected data from this experiment can be used to verify assumptions used in the spatial model and to test the integrated processing of class attributes and spatial descriptions.

The *Select-Rel* test set consists of 173 utterances. The 6 different speakers describe a marked object on a computer screen. They are told to use a spatial relation in their verbal descriptions (see Sec. 4.2 for more detailed information about the experiment).

The word accuracy (WA) of 79.5% is much better than that of the previous data set. However, the feature accuracy (FA) of 87.2% is in the same order of magnitude. In 32 of 173 utterances speech recognition errors caused a complete misinterpretation of the verbal object description. In most of these cases the intended object was interpreted as a reference object. These utterances were ignored in the FEAT_SPEECH tests.

The visual context consists of 6 different scenes that contain only elementary objects, except some unknown out-of-domain objects like in the previous data set. The detection accuracy (DA) of the objects was 64.0%. The recognition errors include 15.1% false type classifications and 9.3% false color classifications.

6.4.1 Verification of the neighborhood assumption

An important assumption used in the spatial model is the introduction of the neighborhood graph:

The speaker will select a reference object in the neighborhood of the intended object when a spatial relation is specified.

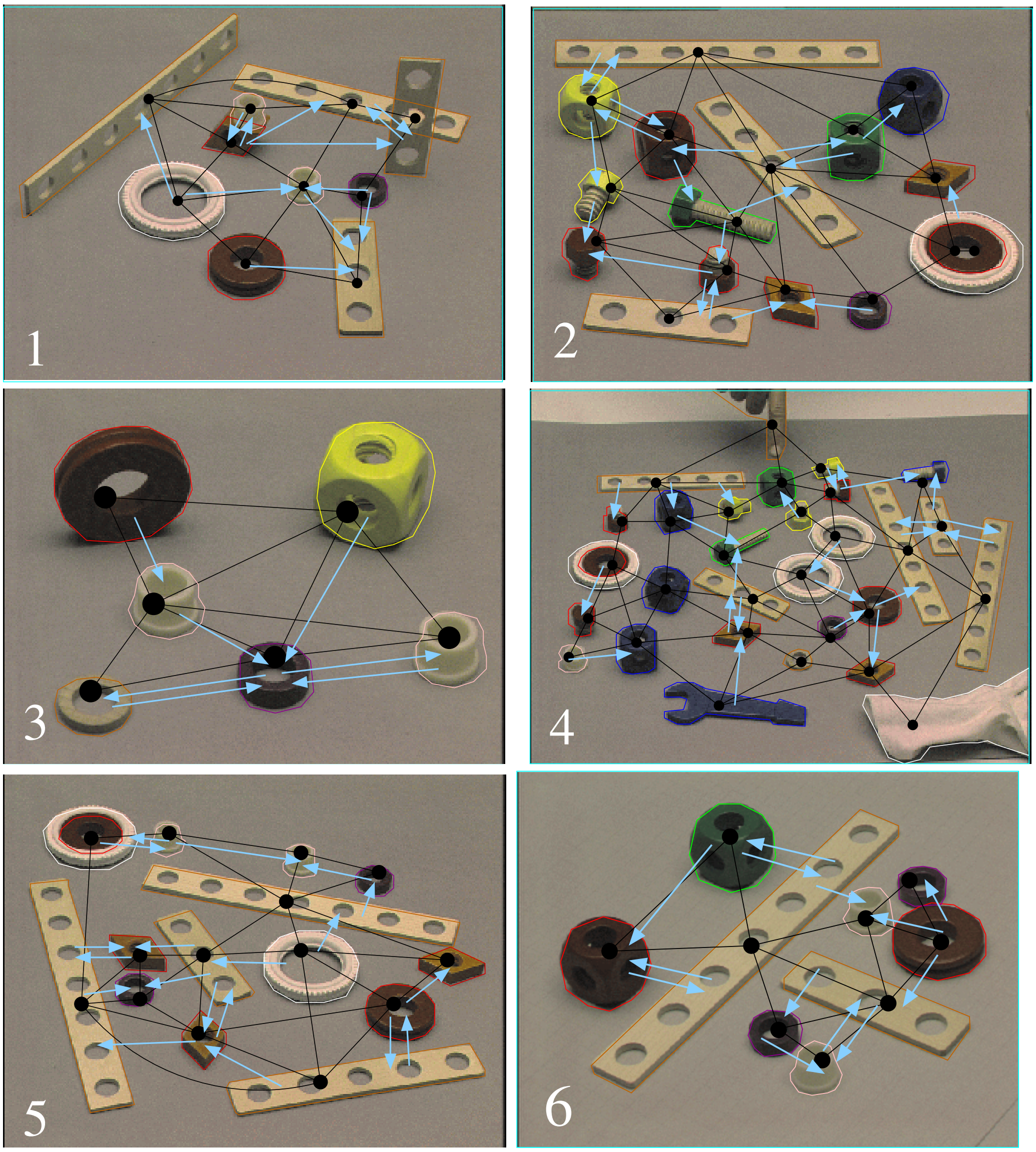


Figure 6.1: Selection of reference objects: The bright arrows point from the reference object to the intended object that were selected by the speaker. The black graph structures represent the calculated neighborhood graphs.

In order to check this assumption all utterances of the test set are examined with regard to the selection of the reference object. The result is shown in Fig. 6.1. The bright arrows are drawn from the selected reference objects to the intended objects. All object pairs are located in the neighborhood of each other.

The black graph structures are calculated based on the neighborhood definition presented in Sec. 4.4.3. The separation threshold is $\Theta_{Sep} = 15\%$. The minimum ratio of the width and the length of the *area in between* is 0.5. If n is the number of objects in the scene each object may be paired with $n-1$ other objects. The computed neighborhood graphs have about $2 \cdot n$ edges. Thus, the average number of reference objects considered is reduced to $\sqrt{n}$. Only 3 out of 174 selected object pairs are not connected in the graph structures (Fig. 6.1: image 1 – two pairs, image 2 – one pair). The main reasons are perspective effects, like the socket that is placed onto the rhomb nut. Nevertheless, the neighborhood definition adequately models the selection of reference objects by the recorded speakers.

6.4.2 Identification results

Before presenting the identification rates, the expectations for this test set will be discussed:

- The speaker shall describe the marked object. Consequently, this object should be included in a ***correct*** system answer which is expected to be near 100%.
- The speakers are explicitly told to specify the type *and* location of the marked object. Therefore, the ***precise*** rate should be high.
- The ***included*** rate is not as relevant as the other rates because the location of the marked object is mostly specified.
- The selection of ***additional*** objects may be caused by an unspecific of abbreviated verbal description, a too unconstrained spatial model, or the use of out-of-domain reference objects.
- A ***false*** identification result may occur because of a not adequate verbal description, a too restricted spatial model, or other un-modeled aspects.
- If verbal features are misrecognized, not detected, or inserted due to speech recognition errors, *the identification rates will decrease*. From a feature accuracy of 87% we expect an average decrease of about 13%.
- If a marked object is misrecognized due a type or color misclassification, *the identification rates will decrease*. 66% of the marked objects are correctly recognized. Consequently the expected decrease of the identification rates is about 34%. The rates will additionally be affected by the detection accuracy (DA 64%) of the other objects because these are used as reference objects.

The identification rates of the integration component are presented in Tab. 6.2. If recognition errors are excluded, 78.6% of the marked objects were correctly selected without any additional objects (***precise*** rate). The identification rate of 93.1% allowing only a unique additional object (+1 ***object***) shows the adequacy of the integration model and, especially, that of the spatial model. Only 4.6% of the marked objects were not selected (***false*** rate).

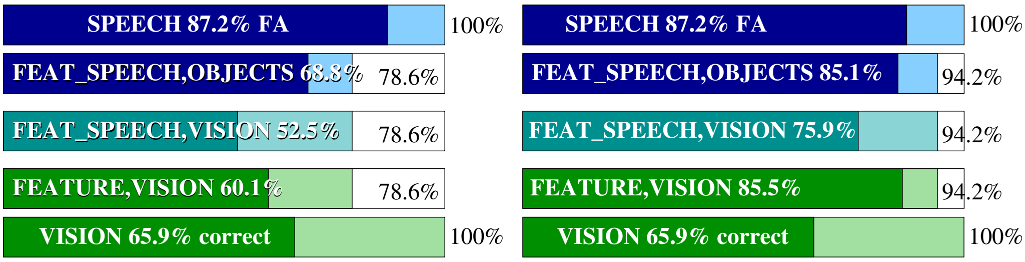
The influence of erroneous input data on the identification rates is visualized in Tab. 6.2(c) and 6.2(d). The base-line results are defined by the FEATURE-OBJECTS

Select-Rel		#utt	precise	included	additional	correct	false
FEATURE	OBJECTS	173	**78.6%**	85.5%	9.8%	95.4%	**4.6%**
FEAT_SPEECH	OBJECTS	141	**68.8%**	76.6%	13.5%	90.1%	9.9%
FEATURE	VISION	173	**60.1%**	70.5%	19.7%	90.2%	9.8%
FEAT_SPEECH	VISION	141	**52.5%**	62.4%	21.3%	83.7%	16.3%

(a) Identification rates

Select-Rel		#utt	precise	+1 object	+2 objects	correct
FEATURE	OBJECTS	173	**78.6%**	**93.1%**	**94.2%**	95.4%
FEAT_SPEECH	OBJECTS	141	**68.8%**	83.0%	**85.1%**	90.1%
FEATURE	VISION	173	**60.1%**	74.6%	**85.5%**	90.2%
FEAT_SPEECH	VISION	141	**52.5%**	64.5%	**75.9%**	83.7%

(b) identification rates with regard to the number of additionally selected objects.



(c) precise

(d) +2 objects

Table 6.2: Identification result for the *Select-Rel* set. The subfigures (c) and (d) visualize identification rates from the table. The feature accuracy is measured on FEAT_SPEECH subset. The *correct* recognition rate is measured on those visual objects that should be identified.

data, they are plotted with light colors. Dark colors indicate that erroneous input data is considered. The decreases of the identification rates are much higher than those of the *Select-Obj* set. For the ***precise*** rates the decrease is in the magnitude of the error rates of the input data: considering SPEECH 12.5%, VISION 24.5%, both 33.2%. If two additionally selected objects are accepted the decrease is substantially reduced. For a single erroneous input channel it is significantly below the input error rates: considering SPEECH 9.7%, VISION 9.2%. If both inputs are noisy their impacts are additively combined resulting in a decrease of 19.4%. The best result is obtained for the ***correct*** identification rate with a decrease of 12.3% (FEAT_SPEECH-VISUAL).

Fig. 6.2 shows how the selection of additional objects is affected by noisy input data. It reveals that the decrease of the *precise* and +1 *object* identification rates is higher for er-

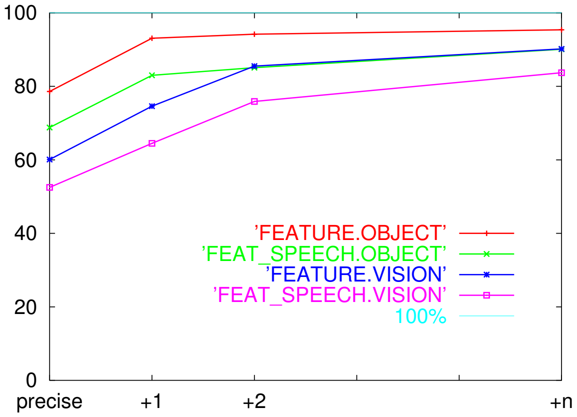


Figure 6.2: Identification rates with regard to the number of selected objects. The '+n' denotes the *correct* identification rates, i.e. arbitrary additional objects may be selected.

roneous *visual* data than for erroneous speech data. This effect disappears if the selection of more than one additional object is accepted.

6.4.3 Qualitative results

The results from the *Select-Rel* test set support the conclusions drawn from the *Select-Obj* results:

- The baufix®domain is adequately modeled by the uncertainty model of the integration component. The neighborhood assumption used in the spatial model is confirmed by the selection of reference objects by the speakers.

- The decrease of the identification rates are *less* than the error rates of the input data. This was confirmed for the ***included***/+2 ***object*** and ***correct*** identification rates.

- The stability of identification rates increases for less constrained system answers. Although ***precise*** results are highly affected by erroneous input data ***correct*** system answers are indeed very stable.

- In the combined case (FEAT_SPEECH,VISION) the impacts of the recognition errors additively combine. Again, both influences seem to be independent.

If the influences from erroneous input data from speech and object recognition are compared a difference can be detected for the *precise* and +1 *object* identification rates:

- The VISION error rate affects the identification rates more than the SPEECH error rate. This may be explained by the experimental constraints because the speaker must refer to *two* objects in the scene in order to identify a single marked object by a spatial relation.

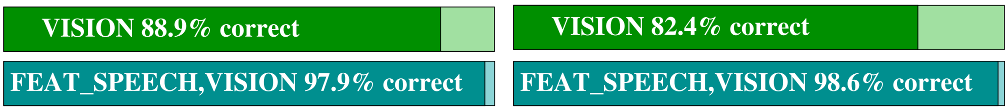
- These effects disappear for less constraint system answers.

Select-Obj		#obj	correct	correct type	correct color
VISION	FEATURE	239	88.3%	90.8%	96.7%
with correction	FEATURE	239	97.9%	97.9%	98.7%
VISION	FEAT_SPEECH	217	**88.9%**	91.7%	95.9%
with correction	FEAT_SPEECH	217	**97.7%**	97.7%	98.2%

(a) recognition rates of precisely selected objects on the *Select-Obj* set.

Select-Rel		#obj	correct	correct type	correct color
VISION	FEATURE	104	85.6%	95.2%	90.4%
with correction	FEATURE	104	99.0%	100.0%	99.0%
VISION	FEAT_SPEECH	74	**82.4%**	91.9%	90.5%
with correction	FEAT_SPEECH	74	**98.6%**	100.0%	98.6%

(b) recognition rates of precisely selected objects on the *Select-Rel* set.



(c) *Select-Obj* set

(d) *Select-Rel* set

Table 6.3: Corrected unique types of precisely selected objects. Subfigures (c) and (d) visualize recognition rates from the tables.

6.5 Object Classification using Speech and Image Features

In the previous section the proposed task 2 (mp-objs) has been quantitatively evaluated. The solution of this task provides the basis for the realization of the other proposed interaction tasks (cf. Sec. 5). An interesting question that can be answered for the previously used test sets in a quantitate manner is related to task 3 (mp-class):

> *Can the object recognition results be improved if features that can be extracted from speech are considered?*

In order to check this question the *precisely* selected objects from the *Select-Obj* and *Select-Rel* data sets are examined. The object recognition rates of these subsets are presented in Tab. 6.3(a) and 6.3(b): between 11.1% and 17.6% of the selected objects have been misrecognized due to type or color misclassifications.

The features extracted from speech and the classification results of the object recognizer are used as evidences in the proposed Bayesian network. The most probable a posteriori state of the *IntendedObjClass* variable is selected as the new object class. Nearly all visual classification errors are corrected considering either FEATURE or FEAT_SPEECH input (see *with correction* in Tab. 6.3).

The restriction of considering only *precise* object selections works as a filter criterion. First, the speaker and the system must insure that both talk about the same scene object. Then the estimation of the most probable object class is performed in a second step.

6.6 Summary

The most basic inference tasks of the integration component proposed in this thesis have been quantitatively evaluated. The task 2 links the visual representation of a scene object with its verbal description that is given by a speaker. Here, a robust system behavior is realized by optionally selecting more than one object if the verbal object description is not precise enough. Erroneous input data, that is caused by speech and object recognition errors, mostly leads to the same system answer or the selection of some additional objects. The intended object still remains in the dialog context and the speaker can select the correct one in the next dialog step by increasing the redundancy of the verbal description.

A comparison of the results with other systems is difficult because of the domain dependency. In section 4.5.1 the previous work by Socher et al. [Soc97, SSP00] was mentioned that this thesis is based on. She evaluated her system QUASI-ACE on the same data sets, but some evaluation conditions were different:

- The *idealized* data is comparable to the FEATURE,OBJECTS input but the identification rates were only counted for 412 of 453 utterances of the *Select-Obj* set and for 98 of 173 utterances of the *Select-Rel* set. In both sets the *false* rates could be reduced from 7.5% (QUASI-ACE) to 4.6% (*Select-Obj*) and from 16.5% (QUASI-ACE) to 4.6% (*Select-Rel*).

- The *text* data is comparable to the FEATURE,VISION input. However, the system QUASI-ACE did not consider *unknown* object regions. Thus the object deletion rate was higher but the object insertion rate was nearly negligible. Here the rate of *not correctly* selected objects (*false+nothing*) could be reduced from 13.6% (QUASI-ACE, 417 utterances) to 9.9% (*Select-Obj*, 453 utterances) and from 21.4% (QUASI-ACE, 84 utterances) to 9.8% (*Select-Rel*, 173 utterances).

- The *speech* data is comparable to the FEAT_SPEECH,VISION input. In the system QUASI-ACE the speech data was processed by a speech understanding component, but only those utterances were considered that were at least partially understood. The rate of the identification categories *false* and *nothing* was reduced from 30.0% (QUASI-ACE, 133 utterances) to 13.0% (*Select-Obj*, 447 utterances) and 24% (QUASI-ACE, 21 utterances) to 16.3% (*Select-Rel*, 141 utterances).

The results of both systems are not directly comparable because of the different evaluation conditions and different evaluation subsets. However, it is shown that the integration component proposed in this thesis models the baufix®domain more adequately and is much more robust if erroneous input data is considered.

Secondly, the realization of task 3 (mp-class) has been evaluated on the *precise* identification results of the *Select-Obj* and *Select-Rel* data sets. The most probable object class is estimated considering the classification results of the vision components and the features extracted from speech input. By this means, a *multi-modal object recognition* scheme is realized. The recognition rate of *precisely* selected objects could be increased up to near 100%.

Chapter 7

Summary and Conclusion

This thesis addresses the problem of relating spoken utterances to the simultaneously perceived visual scene context. The development of systems that integrate verbal and visual information is an extending field of research. It is pushed by various applications like the indexing and querying of video databases, service robotics, augmented reality, document analysis, documentation systems with multi-modal interfaces, or other multi-media systems. Each of these applications has to relate two or more different input modalities.

7.1 The Integration of Speech and Images as a Probabilistic Decoding Process

Speech understanding and vision are the most important abilities in human-human communication, but also the most complex tasks for a machine. A general solution of both tasks is far from being implemented on a computer. It even raises philosophical questions with regard to machine intelligence like the symbol grounding problem and the Chinese room. 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. 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 automatically introduced even for perfect understanding components.

Additionally, the interpretations of the surface modalities are often erroneous or incomplete such that an integrating component must consider noisy and partial interpretations. As a consequence, this thesis treats the correspondence problem as a probabilistic ***decoding process***. This perspective distinguishes this approach from other approaches that propose rule-based translation schemes or integrated knowledge bases and 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. This thesis shows that even though a multi-modal system has to face multiple error sources *the combined signal can be interpreted more stablily than the individual signals*. This has been explicitly shown by a detailed analysis of the identification rates of the implemented system and by the increase of the object recognition rate when features from both modalities are used.

The decoding process is organized as a ***separate subtask*** and is realized by an ***active inference process***. By this means, the integration component becomes independent of the specialized speech understanding and vision components. It is realized as a *principle system component in its own right* like that of Nagel. Different interaction tasks are identified and implemented that can be transferred to any object recognition scheme and any speech understanding component.

7.2 Contributions

The most significant contribution of this work is the demonstration of an ***integration component*** *that robustly combines speech and object recognition results*. The thesis shows that errors occurring in both recognition components can be compensated by combining their interpretation results.

Object recognition results are considered on two different ***layers of abstraction*** in order to increase the robustness of the system. As a consequence, any computational model used in the integration component must be applicable to each layer. For this purpose the 3-d spatial model proposed by Fuhr et al. [FSSS97] was extended and transferred to the 2-d level.

This thesis has successfully applied ***Bayesian networks*** to the task of integrating speech and images. *The correspondence problem has been solved in the language of Bayesian networks in a consistent and efficient way* by using a novel combination of conditioning and elimination techniques. The experimental study has identified Bayesian networks as an adequate formalism for speech and image integrating tasks. The mental models of the speaker are partially reconstructed by estimating conditional probabilities from the data of psycholinguistic experiments. Context dependent shifts of word meanings are modeled by the structure of the network. As an intensional model the inference algorithm is separated from the modeling task. Thus, various inference tasks between the integrated modalities have been formulated using the same integration model. Even questions concerning the internal state of the computational models can be answered like the disambiguation of the reference frame or establishing undetected neighborhood relations.

The proposed Bayesian network scheme for integrating multi-modal input has been applied to a ***construction scenario***. A robot is instructed by a speaker to grasp objects from a table, join them together, and put them down again. In this thesis an integration component is realized that is able to identify objects in the visual scene that are verbally referred to by the speaker. This task is successfully performed despite of vague descriptions, erroneous recognition results, and the use of names with unknown semantics. *Several interaction tasks* have been implemented that perform multi-modal object recognition, link unknown object names to scene objects, disambiguate alternative interpretations of utterances, predict undetected mounting relations, or determine the selected reference frame of the speaker.

7.3 Future Work

The proposed Bayesian network scheme is a very general approach to solve the correspondence problem. It has been successfully applied in a restricted domain raising questions about the ***scalability*** and about the ***portability*** to other domains.

Bayesian networks are an intensional formalism. Thus, the inference algorithm is separated from the modeling task. Consequently, additional aspects and other domains can be modeled without changing the computational framework. The structure of the Bayesian network and the number of exclusive selection variables determine the complexity of the algorithm. A typical number of selection variables for the evaluated data is between one and three with ten to thirty different states. An increased complexity may be introduced by several possible extensions of the system demonstrated so far:

- A serious restriction of the implemented system is the fact that the segmentation results of the vision component cannot be changed on the basis of additionally considered verbal evidences. However, alternative segmentation results may be considered by introducing additional selection variables. The implications on the robustness of the system answers and on the complexity of the inference algorithm are open questions.

- Closely related to this aspect, the system assumes that objects can easily be separated from the background. If this restriction is dropped the number of possible object regions that must be considered drastically increases and the neighborhood assumption must be relaxed.

As mentioned previously the proposed Bayesian network scheme is related to probabilistic graph matching (cf. Sec. 3.4). Thus, a closer examination of the applicability of the proposed algorithm for other weighted graph matching problems will lead to valuable insights. However, the proposed inference algorithm realizes an exact graph matching which is not tractible for a significantly increased problem size. Therefore, the exact inference algorithm must be transformed into an approximate algorithm which seems to be feasible.

Besides the extension of the proposed Bayesian network. There are other serious problems of speech and image integration that are not considered so far.

A prerequisite of the proposed integration scheme is the existence of a finite set of semantically meaningful elementary objects. How can an integration component be designed that considers generic objects that may be defined by geons or other primitives that do not have a semantic meaning?

Only very simple verbal descriptions of complex objects are considered so far. How can computational models be defined that can process more complex structural descriptions like … *consisting of two crossed bars* …?

The last aspect may be described by the duality *modeling* vs. *learning*. What needs to be modeled in a system? What can be learned offline? What can be learned online? The basic competences of the system realized so far are the computational models and the structure of the Bayesian network. The numbers of the conditional probability tables are partially learned offline. The correspondence of names and individual complex objects is learned online. An outlook section proposed an online learning strategy for the names of subassemblies. Bayesian networks may be a good candidate in order to extend the learning abilities of the system because the theory of learning in Bayesian networks is an intensive field of research.

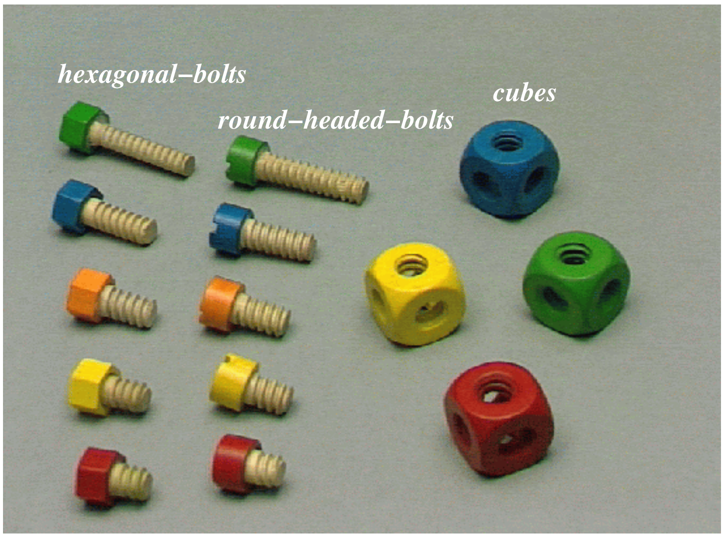
7.4 Final Remarks

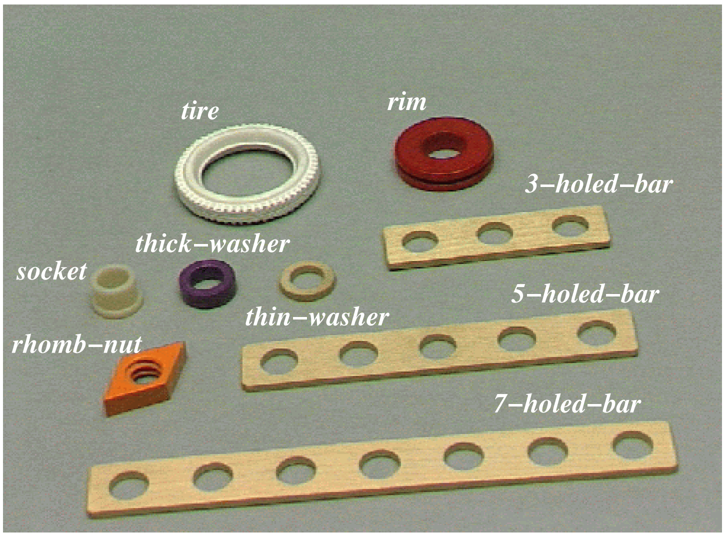
The development of the integration component that is described in this thesis was not straightforward. During the last four years several versions of it were implemented and revised, each step leading to a deeper understanding of the integration task. The final solution is an example that *theory and application can benefit from each other*. The realization of the system lead to new insights in Bayesian networks and vice versa.

Indeed, several lessons have been learned. The combination of two different modalities results in an implicit error correcting strategy. The user obtains robust system answers despite many intermediate results being incorrect. This distinguishes the realized approach form other multi-modal error correction strategies that *ask* the user to give a new input on a different channel. Bayesian networks are an adequate framework for the solution of the correspondence problem. This does not rule out other formalisms like fuzzy logic or Dempster-Shafer. However, the mathematical foundation, intuitive modeling, and ability to learn the parameters of Bayesian networks seem to be promising for future applications.

Appendix A

The elementary objects of baufix®





Bibliography

[ADG84] G. Adorni, M. Di Manzo, and F. Giunchiglia. Natural Language Driven Image Generation. In *Proceedings of COLING*, pages 495–500, 1984.

[AK99] A. Abella and J.R. Kender. From Images to Sentences via Spatial Relations. In *Integration of Speech and Image Understanding: ICCV'99 Workshop*, pages 117–146, Corfu, Greece, 1999. IEEE Computer Society.

[All83] James F. Allen. Maintaining knowledge about temporal intervals. *Communication of the ACM*, 26(11):832–843, 1983.

[All95] James Allen. *Natural language understanding*. Benjamin/Cummings, Redwood City, Calif., 1995.

[AN95] X. Aubert and H. Ney. Large vocabulary continuous speech recognition using word graphs. In *Proc. ICASSP '95*, pages 49–52, Detroit, MI, May 1995.

[AST81] Norihiro Abe, Itsuya Soga, and Saburo Tsuji. A Plot Understanding System on Reference to Both Image and Language. In *International Joint Conference on Artificial Intelligence*, pages 77–84, 1981.

[AWFA85] Steen Andreassen, Marianne Woldbye, Bjorn Falck, and Stig K. Adersen. MUNIN – A Causal Probabilistic Network for Interpretation of Electromyographic Findings. In *International Joint Conference on Artificial Intelligence*, pages 366–372, 1985.

[BB82] Dana Ballard and Chris Brown. *Computer Vision*. Englewood Cliffs, NJ:Prentice-Hall, 1982.

[BBW92] Michael K. Brown, Bruce M. Buntschuh, and Jay G. Wilpon. SAM: A Perceptive Spoken Language Understanding Robot. *IEEE Transactions on Systems, Man, and Cybernatics*, 22(6):1390–1402, 1992.

[BFF$^+$01] C. Bauckhage, G.A. Fink, J. Fritsch, F. Kummert, F. Lömker, G. Sagerer, and S. Wachsmuth. An Integrated System for Cooperative Man-Machine Interaction. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Banff, Canada, 2001. to appear.

[BFKS99] C. Bauckhage, J. Fritsch, F. Kummert, and G. Sagerer. Towards a Vision System for Supervising Assembly Processes. In *Proc. Symposium on Intelligent Robotic Systems (SIRS'99)*, pages 89–98, 1999.

[Bin71] T.O. Binford. Visual perception by computer. In *IEEE conference on Systems and Control*, Miami, December 1971.

[BJ72] John D. Bransford and Marcia K. Johnson. Context Prerequisites for Understanding: Some Investigations of Comprehension and Recall. *Journal of Verbal Learning and Verbal Behaviour*, 11:717–726, 1972.

[BJPW95] Christel Brindöpke, Michaela Johanntokrax, Arno Pahde, and Britta Wrede. "'Darf ich dich Marvin nennen?"' — Instruktionsdialoge in einem Wizard-of-Oz-Szenario: Materialband. Report 7/95, Sonderforschungsbereich 360 "'Situierte Künstliche Kommunikatoren"', Universität Bielefeld, 1995.

[BKS98] C. Bauckhage, F. Kummert, and G. Sagerer. Modeling and Recognition of Assembled Objects. In *Proc. Annual Conference of the IEEE Industrial Electronics Society (IECON'98)*, pages 2051–2056, 1998.

[BLM$^+$98] Tom Brondsted, Lars Bo Larsen, Michael Manthey, Paul McKevitt, Thomas Moeslund, and Kristian G. Olesen. The Intellimedia Workbench – a Generic Environment for Multimodal Systems. In *International Conference on Spoken Language Processing*, pages 273–276, 1998.

[BP99] Hans Brandt-Pook. *Eine Sprachverstehenskomponente in einem Konstruktionsszenario*. PhD thesis, Bielefeld University, 1999.

[BPFWS99] Hans Brandt-Pook, Gernot A. Fink, Sven Wachsmuth, and Gerhard Sagerer. Integrated recognition and interpretation of speech for a construction task domain. In *Proc. of the International Conference on Human Computer Interaction (HCI)*, volume 1, pages 550–554, 1999.

[CFH97] Eliseo Clementini, Paolino Di Felice, and Daniel Hernández. Qualitative representation of positional information. *Artificial Intelligence*, 95:317–356, Sep. 1997.

[Cla73] H.H. Clark. Space, time, semantics, and the child. In T.E. Moore, editor, *Cognitive developement and the aquisition of language*, pages 65–110. Academic Press, New York, 1973.

[Col94] Robin Collier. An Historical Overview of Natural Language Systems that Learn. *Artificial Intelligence Review*, 8:17–54, 1994.

[CR97] Tsuhan Chen and Ram R. Rao. Audio-Visual Interaction in Multimedia Communication. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 179–182, 1997.

[CS73] L.A. Cooper and R.N. Shepard. Chronometric studies of the rotation of mental images. In W.G. Chase, editor, *Visual information processing*. Academic Press, New York, 1973.

[CS95] Rajiv Chopra and Rohini K. Srihari. Control Structures for Incorporating Picture-Specific Contex in Image Interpretation. In *International Joint Conference on Artificial Intelligence*, pages 50–55, 1995.

[Dec96] R. Dechter. Topological parameters for time-space tradeoffs. In *Uncertainty in Artificial Intelligence (UAI-96)*, pages 220–227, 1996.

[Dec97] Rina Dechter. Mini-Buckets: A General Scheme For Generating Approximations In Automated Reasoning. In *International Joint Conference on Artificial Intelligence*, 1997.

[Dec98] Rina Dechter. Bucket elimination: a unifying framework for probabilistic inference. In Michael I. Jordan, editor, *Learning in graphical models*. Kluwer Academic Publisher, Dordecht, The Netherlands, 1998.

[DHN76] R. Duda, P.E. Hart, and N.J. Nilsson. Subjective Bayesian methods for rule-based inference systems. In *Proc. Natl. Comp. Conf. (AFIPS)*, volume 45, pages 1075–1082, 1976.

[Dic99] Sven J. Dickinson. Object Representation and Recognition. In E. Lepore and Z. Pylyshyn, editors, *Rudgers University Lectures on Cognitive Science*, pages 172–207. Basil Blackwell publishers, 1999.

[DPR92] Sven J. Dickinson, Alex P. Pentland, and Azriel Rosenfeld. 3-D Shape Recovery Using Distributed Aspect Matching. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 14(2):174–198, 1992.

[DR96] R. Dechter and I. Rish. Guess or think? Hybrid algorithms for SAT. In *Principles and Practice of Constraint Programming (CP-96)*, 1996.

[EFD96] Y. El-Fattha and R. Dechter. An evaluation of structural parameters for probabilistic reasoning. In *Uncertainty in Artificial Intelligence (UAI-96)*, pages 244–251, 1996.

[Fau93] Olivier D. Faugeras. *Three-dimensional computer vision*. MIT Press, 1993.

[FHRG94] C. Fisher, G. Hall, S. Rakovitz, and L. Gleitman. When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92(1):333–375, 1994.

[Fin99] Gernot A. Fink. Developing HMM-based recognizers with ESMERALDA. In Václav Matoušek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin, 1999. Springer.

[FLGD87] G.W. Furnas, T.K. Landauer, L.M. Gomez, and S.T. Dumais. The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, 30(11):964–971, 1987.

[FSSS97] Thomas Fuhr, Gudrun Socher, Christian Scheering, and Gerhard Sagerer. A three-dimensional spatial model for the interpretation of image data. In P. Olivier and K.-P. Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 103–118. Lawrence Erlbaum Associates, 1997.

[Gap94] Klaus-Peter Gapp. Basic meanings of spatial relations: Computation and evaluation in 3d space. In *Proc. of AAAI-94*, pages 1393–1398, Seattle, WA, 1994.

[GLP84] W. Grimson and T. Lozano-Pérez. Model-based recognition and localization from sparse range or tactile data. *International Journal of Robotics Research*, 3(3):3–35, 1984.

[GZ92] David Goddeau and Vistor Zue. Integrating probabilistic lr-parsing into speech understanding systems. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 181–184, 1992.

[Har90] Steven Harnad. The Symbol Grounding Problem. *Physica D*, 42:335–346, 1990.

[HB00] R.J. Howarth and H. Buxton. Conceptual description from monitoring and watching image sequences. *Image and Vision Computing*, 18:105–135, 2000.

[Her86] Annette Herskovits. *Language and Spatial Cognition*. Cambridge University Press, 1986.

[Her90] Theo Herrmann. Vor, hinter, rechts und links: das 6H-Modell. *Zeitschrift für Literaturwissenschaft und Linguistik*, 78:117–140, 1990.

[Her94] Daniel Hernández. *Qualitative Representation of Spatial Knowledge*, volume 804 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, 1994.

[HKM$^+$94] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and & J. Weber. Automatic symbolic traffic scene analysis using belief networks. In *Proc. AAAI-94*, pages 966–972, 1994.

[HKN91] N. Heinze, W. Krüger, and H.-H. Nagel. Berechnung von Bewegungsverben zur Beschreibung von aus Bildfolgen gewonnenen Fahrzeugtrajektorien in Strassenverkehrsszenen. *Informatik – Forschung und Entwicklung*, 6:51–61, 1991.

[HN00] M. Haag and H.-H. Nagel. Incremental recognition of traffic situations from video image sequences. *Image and Vision Computing*, 18:137–153, 2000.

[HR95] G. Herzog and K. Rohr. Integrating Vision and Language: Towards Automatic Description of human movements. In *Proceedings of the 19th Annual German Conference on Artificial Intelligence (KI-95)*, 1995.

[HW94] A. Hauenstein and H. Weber. An investigation of tightly coupled time synchronous speech language interface using a unification grammar. In *Proceedings of the Workshop on Integration of Natural Language and Speech Processing at AAAI 94*, pages 42–49, 1994.

[IHTS98] Ichiro Ide, Reiko Hamada, Hidehiko Tanaka, and Shuichi Sakai. News Video Classification based on Semantic Attributes of Captions. In *Proceedings 6th ACM Intl. Multimedia Conference*, pages 60–61, 1998.

[Int99] Stephen Sean Intille. *Visual Recognition of Multi-Agent Action*. PhD thesis, Massachusetts Institute of Technology, 1999.

[IT98] Ichiro Ide and Hidehiko Tanaka. Automatic Semantic Analysis of Television News Captions. In *Proceedings 3rd Intl. Workshop on Information Retrieval with Asian Languages*, 1998.

[Jac85] Ray S. Jackendoff. *Semantics and Cognition*. MIT Press, 1985.

[Jac87] Ray Jackendoff. On Beyond Zebra: The relation of linguistic and visual information. *Cognition*, 26:89–114, 1987.

[Jac89] Ray S. Jackendoff. *Consciousness and the computational mind*. MIT Press, 1989.

[Jen96] Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press Limited, London, 1996.

[KFSB98] F. Kummert, G.A. Fink, G. Sagerer, and E. Braun. Hybrid Object Recognition in Image Sequences. In *Proc. International Conference on Pattern Recognition (ICPR'98)*, volume II, pages 1165–1170, 1998.

[KK99] S. Kronenberg and F. Kummert. Structural dependencies between syntactic relations: A robust parsing model for extrapositions in spontaneous speech. In *ASRU: IEEE International Workshop on Automatic Speech Recognition and Understanding*, 1999. (http://asru99.research.att.com).

[KNPS93] F. Kummert, H. Niemann, R. Prechtel, and G. Sagerer. Control and Explanation in a Signal Understanding Environment. *Signal Processing, special issue on 'Intelligent Systems for Signal and Image Understanding'*, 32:111–145, 1993.

[Kos94] Stephen M. Kosslyn. *Image and Brain*. MIT Press, 1994.

[Kro01] Susanne Kronenberg. *Cooperation in Human-Computer Communication*. PhD thesis, Bielefeld University, 2001.

[KWK99] Susanne Kronenberg, Sven Wachsmuth, and Franz Kummert. Disambiguation of utterances by visual context information. In *Mustererkennung 99, 21. DAGM-Symposium Bonn*, pages 338–347, Berlin, 1999. Springer-Verlag.

[Lee89] Kai-Fu Lee. *Automatic Speech Recognition*. Kluwer Academic Publishers, 1989.

[LGS86] J.D. Lowrance, T.D. Garvey, and T.M. Strat. A framework for evidential reasoning systems. In *Proc. 5th Matl. Conf. on AI (AAAI-86)*, pages 896–901, Philadelphia, 1986.

[LJ93] Barbara Landau and Ray Jackendoff. “What” and “where” in spatial language and spatial cognition. *Behavioural and Brain Sciences*, 16:217–265, 1993.

[LP74] Elizabeth F. Loftus and John C. Palmer. Reconstruction of Automobile Destruction: An Example of the Interaction Between Language and Memory. *Journal of Verbal Learning and Verbal Behaviour*, 13:585–589, 1974.

[LVW98] Mihai Lazarescu, Svetha Venkatesh, and Geoff West. Combining NL processing and video data to query American Football. In *International Conference on Pattern Recognition*, pages 1238–1240, 1998.

[Mar82] David Marr. *Vision*. W.H. Freeman, 1982.

[Mar96] Diego Marconi. Work on the Integration of Language and Vision at the University of Torino. In *Artificial Intelligence Review*, volume 10, pages 15–20. Kluwer Academic Publishers, Netherlands, 1996.

[May93] T. Maybury, editor. *Intelligent MultiMedia Interfaces*. AAAI Press/The MIT Press, 1993.

[McC77] J. McCarthy. Epistemological Problems in Artificial Intelligence. In *International Joint Conference on Artificial Intelligence*, Cambrigde, Massachusetts, 1977.

[Mes95] B. Messmer. *Efficient Graph Matching Algorithms for Preprocessed Model Graphs*. PhD thesis, Institut für Informatik und angewandte Mathematik, Universität Bern, Switzerland, 1995.

[MM76] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[MP72] M. Minski and S. Papert. *Perceptrons*. MIT Press, 1972.

[MP96] Christina Meini and Alfredo Paternoster. Understanding Language through Vision. In *Artificial Intelligence Review*, volume 10, pages 37–48. Kluwer Academic Publishers, Netherlands, 1996.

[MRB00] Helmut Mangold and Peter Regel-Brietzmann. Usability von Spracherkennungssystemen als Voraussetzung für eine breite Anwendung. In *KONVENS 2000 Sprachkommunikation*, number 161 in ITG-Fachbericht. VDE Verlag, 2000.

[Muk97] Amitabha Mukerjee. Neat vs. Scruffy: A survey of Computational Models for Spatial Expressions. In P. Olivier and Klaus-Peter Gapp, editors, *Representation and Processing of Spatial Expressions*. Lawrence Erlbaum Associates, 1997.

[MWH93] W. Maaß, P. Wazinski, and G. Herzog. VITRA GUIDE: Multimodal route descriptions for computer assisted vehicle navigation. In *Sixth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems*, pages 104–112, June 1993.

[Nag94] Hans-Hellmut Nagel. A Vision of 'Vision and Language' Comprises Action: An Example from Road Traffic. *Artificial Intelligence Review*, 8:189–214, 1994.

[Nag99] Hans-Hellmut Nagel. From Video to Language–A Detour via Logic vs. Jumping to Conclusions. In Sven Wachsmuth and Gerhard Sagerer, editors, *Integration of Speech and Image Understanding*, pages 79–100, Corfu, Greece, 1999. IEEE Computer Society.

[NB90] G.S. Novak and W.C. Bulko. Understanding Natural Language with Diagrams. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 465–470, 1990.

[NKFH98] M.R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang. Probabilistic Multimedia Objects (Multijects): A Novel Approach to Video Indexing and Retrieval in Multimedia Systems. In *International Conference on Image Processing*, pages 536–540, 1998.

[NN83] B. Neumann and H. Nova. Event Models for Recognition and Natural Language Description of Events in Real-World Image Sequences. In *Proceedings of IJCAI*, pages 724–726, 1983.

[NO99] Hermann Ney and Stefan Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, 16(5):64–83, 1999.

[NR95] Katashi Nagao and Jun Rekimoto. Ubiquitous talker: Spoken language interaction with real world objects. In *International Joint Conference on Artificial Intelligence*, pages 1284–1290, 1995.

[NSF$^+$95] Uta Neave, Gudrun Socher, Gernot A. Fink, Franz Kummert, and Gerhard Sagerer. Generation of Language Models Using the Results of Image Analysis. In *Proceedings of EUROSPEECH'95*, 1995.

[NSS58] A. Newell, J. C. Shaw, and H. A. Simon. The elements of a theory of human problem solving. *Psychological Review*, 65:151–166, 1958.

[OMiT94] Patrick Olivier, Toshiyuki Maeda, and Jun ichi Tsuji. Automatic depiction of spatial description. In *AAAI-94*, pages 1405–1410, 1994.

[O'S00] Douglas O'Shaughnessy. *Speech Communications: Human and Machine*. Addison-Wesley, Reading, Massachusetts, 2 edition, 2000.

[Pai71] A. Paivio. *Imagery and Verbal Processes*. Holt Rinehart & Winston, New York, 1971.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, California, 1988.

[Pin84] Steven Pinker. Visual cognition: An introduction. *Cognition*, 18:1–63, 1984.

[Pin89] S. Pinker. *Learnability and Cognition*. MIT Press, 1989.

[Raj94] R. Rajagopalan. A Model for Integrated Qualitative Spatial and Dynamic Reasoning about Physical Systems. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 1411–1417, 1994.

[RB94] Raymond D. Rimey and Christopher M. Brown. Control of Selective Perception Using Bayes Nets and Decision Theory. *International Journal of Computer Vision*, 12(2/3):173–207, 1994.

[RJ93] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[RM87] R. Reiter and A.K. Mackworth. A Logical Framework for Depiction and Image Interpretation. Technical report, University of British Columbia, 1987.

[RN95] Stuart Russell and Peter Norvig. *Artificial Intelligence: a modern approach*. Prentice Hall, 1995.

[RP98a] Deb Roy and Alex Pentland. Learning Words from Natural Audio-Visual Input. In *International Conference on Spoken Language Processing*, volume 4, pages 1279–1282, 1998.

[RP98b] Deb Roy and Alex Pentland. Word Learing in a Multimodal Environment. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE Press, 1998.

[RS88] G. Retz-Schmidt. Various views on spatial prepositions. *AI Magazine*, 9(2):95–105, 1988.

[RSP99] Deb Roy, Bernt Schiele, and Alex Pentland. Learning Audio-Visual Associations using Mutual Information. In *Integration of Speech and Image Understanding*, pages 147–163. IEEE Press, 1999.

[SB94] R.K. Srihari and D.T. Burhans. Visual Semantics: Extracting Visual Information from Text Accompanying Pictures. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 793–798, 1994.

[SC82] R.N. Shepard and L.A. Cooper. *Mental images and their transformations*. MIT Press, 1982.

[Sch96] Jürgen Schürmann. *Pattern Classification*. Wiley, 1996.

[Sea80] J. Searle. Minds, brains, and programs. *Behavioral & Brain Sciences*, 3:417–458, 1980.

[Sho76] E.H. Shortliffe. *Computer-based medical consultation: MYCIN*. Elsevier Science, Amsterdam, 1976.

[Sis96] Jeffrey M. Siskind. A Computational Study of Cross-Situation Techniques for Learning Word-to-Meaning Mappings. *Cognition*, 61:39–91, 1996.

[Sis98] Jeffrey M. Siskind. Visual Event Perception. In *Proceedings of the Ninth NEC Research Symposium*, 1998.

[SMW96] Bernhard Suhm, Brad Myers, and Alex Waibel. Interactive Recovery from Speech Recognition Errors in Speech User Interfaces. In *International Conference on Spoken Language Processing*, pages 865–868, 1996.

[Soc97] Gudrun Socher. *Qualitative Scene Descriptions from Images for Integrated Speech and Image Understanding*, volume DISKI 170 of *Dissertationen zur Künstlichen Intelligenz*. infix-Verlag, Sankt Augustin, 1997.

[Sri94] Rohini K. Srihari. Computational Models for Integrating Linguistic and Visual Information: A Survey. *Artificial Intelligence Review*, 8:349–369, 1994.

[Sri95] Rohini K. Srihari. Use of Collateral Text in Understanding Photos. *Artificial Intelligence Review*, 8:409–430, 1995.

[SSP00] G. Socher, G. Sagerer, and P. Perona. Bayesian reasoning on qualitative descriptions from images and speech. *Image and Vision Computing*, 18:155–172, 2000.

[SSS$^+$97] Shin'ichi Satoh, Toshio Sato, Michael A. Smith, Yuichi Nakamura, and Takeo Kanade. Name-It: Naming and Detecting Faces in News Video. In *International Joint Conference on Artificial Intelligence*, pages 1488–1493, 1997.

[SWBPK99] Gerhard Sagerer, Sven Wachsmuth, Hans Brandt-Pook, and Franz Kummert. Ein Raummodell für die Benennung von Objekten in 3D-Szenen. In Gerd Rickeit, editor, *Richtungen im Raum*, pages 203–232. Deutscher Universitätsverlag, 1999.

[TNKS98] Takuya Takahashi, Satoru Nakanishi, Yoshinori Kuno, and Yoshiaki Shirai. Helping Computer Vision by Verbal and Nonverbal Communication. In *International Conference on Pattern Recognition*, pages 1216–1218, 1998.

[TR87] S. Truve and W. Richards. From Waltz to Winston (via the Connection Table). In *International Conference on Computer Vision*, pages 393–404, 1987.

[TVD$^+$97] J.K. Tsotsos, G. Verghese, S. Dickinson, M. Jenkin, E. Milios, F. Nuflo, S. Stevenson, M. Black, D. Metaxas, S. Culhane, Y. Ye, and R. Mann. PLAYBOT: A Visually-Guided Robot for Physically Disabled Children. *Image and Vision Computing*, 1997.

[UM82] Leslie G. Ungerleider and Mortimer Mishkin. Two Cortical Visual Systems. In David J. Ingle, Melvyn A. Goodale, and Richard J.W. Mansfield, editors, *Analysis of Visual Behaviour*, pages 549–586. MIT Press, 1982.

[Vor01a] Constanze Vorwerg. Kategorisierung von Grössen- und Formattributen. In A. Zimmer, K. Lange, K. Bäuml, R. Loose, R. Scheuchenfpflug, O. Tucha, H. Schnell, and R. Findl, editors, *Abstracts der 43. Tagung experimentell arbeitender Psychologen*, Experimentelle Psychologie. Lengerich: Pabst Science Publishers, 2001.

[Vor01b] Constanze Vorwerg. Kategorisierung von Grössen- und Formattributen. In *Posterbeitrag auf der 43. Tagung experimentell arbeitender Psychologen*, Regensburg, Apr. 9–11 2001.

[Vos91] George Vosselman. *Relational Matching*. Number 628 in Lecture Notes in Computer Science. Springer-Verlag, 1991.

[VSF$^+$97] Constanze Vorwerg, Gudrun Socher, Thomas Fuhr, Gerhard Sagerer, and Gert Rickheit. Projective relations for 3D space: computational model, application, and psychological evaluation. In *Proceedings of the 14th National Joint Conference on Artificial Intelligence AAAI-97*, pages 159–164, Rhode Island, 1997.

[Wal81] David L. Waltz. Generating and understnading scene descriptions. In B. Webber and I. Sag, editors, *Elements of Discourse Understanding*, pages 266–282. Cambridge Univ. Press, 1981.

[WAWB$^+$94] M. Woszczyna, N. Aoki-Waibel, F.D. Buo, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavic, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz,

B. Suhm, M. Tomita, and A. Waibel. JANUS 93: Towards Spontaneous Speech Translation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 345–348, 1994.

[WBPK^{+}99] Sven Wachsmuth, Hans Brandt-Pook, Franz Kummert, Gudrun Socher, and Gerhard Sagerer. Integration of vision and speech understanding using bayesian networks. *Videre: A Journal of Computer Vision Research*, 1(4):62–83, 1999.

[WBPS^{+}99] Sven Wachsmuth, Hans Brandt-Pook, Gudrun Socher, Franz Kummert, and Gerhard Sagerer. Multilevel integration of vision and speech understanding using bayesian networks. In Hendrik I. Christensen, editor, *Computer Vision Systems: First International Conference*, volume 1542 of *Lecture Notes in Computer Science*, pages 231–254, Las Palmas, Gran Canaria, Spain, January 1999. Springer-Verlag.

[WFKS00] Sven Wachsmuth, Gernot A. Fink, Franz Kummert, and Gerhard Sagerer. Using speech in visual object recognition. In G. Sommer, N. Krüger, and C. Perwass, editors, *Mustererkennung 2000, 22. DAGM-Symposium Kiel*, Informatik Aktuell, pages 428–435. Springer, 2000.

[WFS98] Sven Wachsmuth, Gernot A. Fink, and Gerhard Sagerer. Integration of parsing and incremental speech recognition. In *Proceedings of the European Signal Processing Conference (EUSIPCO-98)*, volume 1, pages 371–375, Rhodes, September 1998.

[Win73] Terry Winograd. A Procedural Model of Language Understanding. In Roger C. Schank, editor, *Computer Models of Thought and Language*, pages 152–186. Freeman, San Francisco, Calif., 1973.

[WKt^{+}99] W.A.J.J. Wiegerinck, H.J. Kappen, E.W.M.T. ter Braak, W.J.P.P. ter Burg, M.J. Nijman, Y.L. O, and J.P. Neijt. Approximate inference for medical diagnosis. *Pattern Recognition Letters*, 20:1231–1239, 1999.

[WS99] Sven Wachsmuth and Gerhard Sagerer. Connecting concepts in vision and speech processing. In *Integration of Speech and Image Understanding : ICCV'99 Workshop*, pages 1–20. IEEE Computer Society, 1999.

[WSVY97] Alex Waibel, Bernhard Suhm, Minh Tue Vo, and Jiee Yang. Multimodal Interfaces for Multimedia Information Agents. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 167–170, 1997.

[YTK84] M. Yokota, R. Taniguchi, and E. Kawaguchi. Language-Picture Question-Answering Through Common Semantic Representation and its Application to the World of Weather Report. In Leonard Bolc, editor, *Natural Language Communication with Pictorial Information Systems*, pages 203–254. Springer-Verlag, 1984.

[ZV88] U. Zernik and B.J. Vivier. How Near Is Too Far? Talking about Visual Images. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society*, pages 202–208. Lawrence Erlbaum Associates, 1988.

Index