

Contents

1	Introduction	13
2	Problem Statement	17
2.1	Robust Processing in Human-Computer Interaction	17
2.2	Basic Principles of Computer Vision	19
2.3	Basic Principles of Automatic Speech Understanding	23
2.4	Integration of Speech and Image Processing	
	– An Overview	27
2.4.1	Psychological experiments and the level of information processing	27
2.4.2	Linguistics and the symbol grounding problem	28
2.4.3	Spatial cognition	29
2.4.4	A categorization of computational systems	29
2.5	The Correspondence Problem	32
2.5.1	Knowledge representation and control structures	35
2.5.2	Spatial models	40
2.5.3	Learning	43
2.6	Other Related Work	50
2.7	Contributions	53
2.7.1	A probabilistic translation scheme	53
2.7.2	A separate integration and interaction component for speech un- derstanding and vision base-line systems	53
2.7.3	The choice of the application area	54
2.7.4	Inference and learning	54
3	A Model for Uncertainty	57
3.1	Intensional and extensional models	58
3.2	Bayesian Networks	60
3.2.1	Definition of Bayesian networks	62
3.2.2	Modeling in Bayesian networks	63
3.2.3	How to get those numbers? Some simplification	65
3.2.4	Modeling corresponding variables	68
3.3	Inference in Bayesian Networks	71
3.3.1	I-maps, moral graphs, and d-separation	72

3.3.2	Singly connected networks	74
3.3.3	Coping with loops	75
3.3.4	A conditional bucket elimination scheme	84
3.4	Relation to Graph Matching	91
3.5	Applications of Bayesian Networks	93
3.6	Bayesian networks for integration of speech and images	99
3.6.1	Modeling principles	99
3.6.2	Inference methods	101
3.6.3	An application to human-computer interaction	102
4	Modeling	103
4.1	Scenario and Domain Description	103
4.2	Experimental Data	105
4.3	The General System Architecture	107
4.3.1	The speech understanding and dialog components	108
4.3.2	The object recognition component	111
4.3.3	Speech understanding and vision results	113
4.4	Spatial Modeling	117
4.4.1	A model for 3-d projective relations	117
4.4.2	The spatial model in two dimensions	118
4.4.3	The neighborhood graph	122
4.4.4	Localization attributes	124
4.4.5	Summary	126
4.5	Object Identification using Bayesian Networks	128
4.5.1	Previous work	128
4.5.2	Starting points for improvements	131
4.5.3	An extended Bayesian model for object classes	133
4.5.4	A Bayesian model for spatial relations	137
4.5.5	Modeling structural relationships	141
4.5.6	Integrating the what and where	142
4.6	Summary	144
5	Inference and Learning	147
5.1	Establishing referential links	147
5.2	Interaction of speech and image understanding	150
5.2.1	The most probable class of the intended object	150
5.2.2	Interpretation of structural descriptions	156
5.2.3	Unknown object names	159
5.2.4	Disambiguating alternative interpretations of an utterance	159
5.2.5	Disambiguating the selected reference frame	163
5.2.6	Detection of neighborhood relations	164
5.3	Further Learning Capabilities	166

6	Results	169
6.1	Test Sets	169
6.2	Classification of System Answers	170
6.3	Results on the <i>Select-Obj</i> test set	171
6.4	Results on the <i>Select-Rel</i> test set	173
6.4.1	Verification of the neighborhood assumption	173
6.4.2	Identification results	175
6.4.3	Qualitative results	177
6.5	Object Classification using Speech and Image Features	178
6.6	Summary	179
7	Summary and Conclusion	181
7.1	The Integration of Speech and Images as a Probabilistic Decoding Process	181
7.2	Contributions	182
7.3	Future Work	183
7.4	Final Remarks	184
A	The elementary objects of <i>baufix</i>[®]	185