

Contents

1 Introduction	1
1.1 Focus of Attention Tracking in Meetings	3
1.2 Approach	7
1.3 Outline	10
2 Background and Related Work	13
2.1 Human Attention	13
2.1.1 Computational models of attention	15
2.2 Where We Look Is Where We Attend To	16
2.3 Gaze and Attention During Social Interaction	19
2.4 Cues for the Perception of Gaze	21
2.5 Eye Gaze Tracking Techniques	22
2.6 Head Pose Tracking	24
2.6.1 Vision-Based Methods	25
2.7 Summary	27
3 Detecting and Tracking Faces	29
3.1 Appearance Based Face Detection	29
3.2 Face Detection Using Color	30
3.3 A Stochastic Skin-Color Model	31

3.4	Locating Faces Using the Skin-Color Model	32
3.5	Tracking Faces With an Omni-Directional Camera	33
3.5.1	Discussion	36
4	Head Pose Estimation Using Neural Networks	37
4.1	Data Collection	38
4.1.1	Data Collection With a Pan-Tilt-Zoom Camera	39
4.1.2	Data Collection With the Omni-Directional Camera	40
4.2	Image Preprocessing	41
4.2.1	Histogram Normalization	41
4.2.2	Edge Detection	42
4.3	Neural Network Architecture	43
4.4	Other Network Architectures	44
4.5	Experiments and Results With Pan-Tilt-Zoom Camera Images	45
4.5.1	Error Analysis	46
4.5.2	Generalization to Different Illumination	48
4.5.3	A Control-Experiment to Show the Usefulness of Edge Features	50
4.6	Experiments and Results With Images From the Omni-Directional Camera	52
4.6.1	Adding Artificial Training Data	53
4.6.2	Comparison	53
5	From Head Orientation to Focus of Attention	55
5.1	Unsupervised Adaptation of Model Parameters	58
5.2	Experimental Results	59
5.2.1	Meetings With Four Participants	63
5.2.2	Meetings With Five Participants	64
5.2.3	Upper Performance Limits Given Neural Network Outputs	64
5.3	Panoramic Images Versus High-Resolution Images	66
5.4	Summary	67

6	Head Pose versus Eye-Gaze	69
6.1	Data Collection	69
6.2	Contribution of Head Orientation to Gaze	70
6.3	Predicting the Gaze Target Based on Head Orientation	73
6.3.1	Labeling Based on Gaze Direction	73
6.3.2	Prediction Results	74
6.4	Discussion	75
7	Combining Pose Tracking with Likely Targets of Attention	77
7.1	Predicting Focus Based on Sound	78
7.1.1	Sound-Only Based Prediction Results	80
7.2	Combining Head Pose and Sound to Predict Focus	80
7.3	Using Temporal Speaker Information to Predict Focus	81
7.3.1	Experimental Results	84
7.3.2	Combined Prediction Results	85
7.4	Summary	86
8	Portability	89
8.1	Data Collection at CMU	90
8.2	Head Pan Estimation Experiments	90
8.2.1	Training New Networks from Scratch	91
8.2.2	Adapting a Trained Network	92
8.3	Focus of Attention Detection Results	95
8.4	Discussion	97
9	Focus of Attention in Context-Aware Multimodal Interaction	99
10	Conclusions	103
10.1	Future Work	105
	Bibliography	107